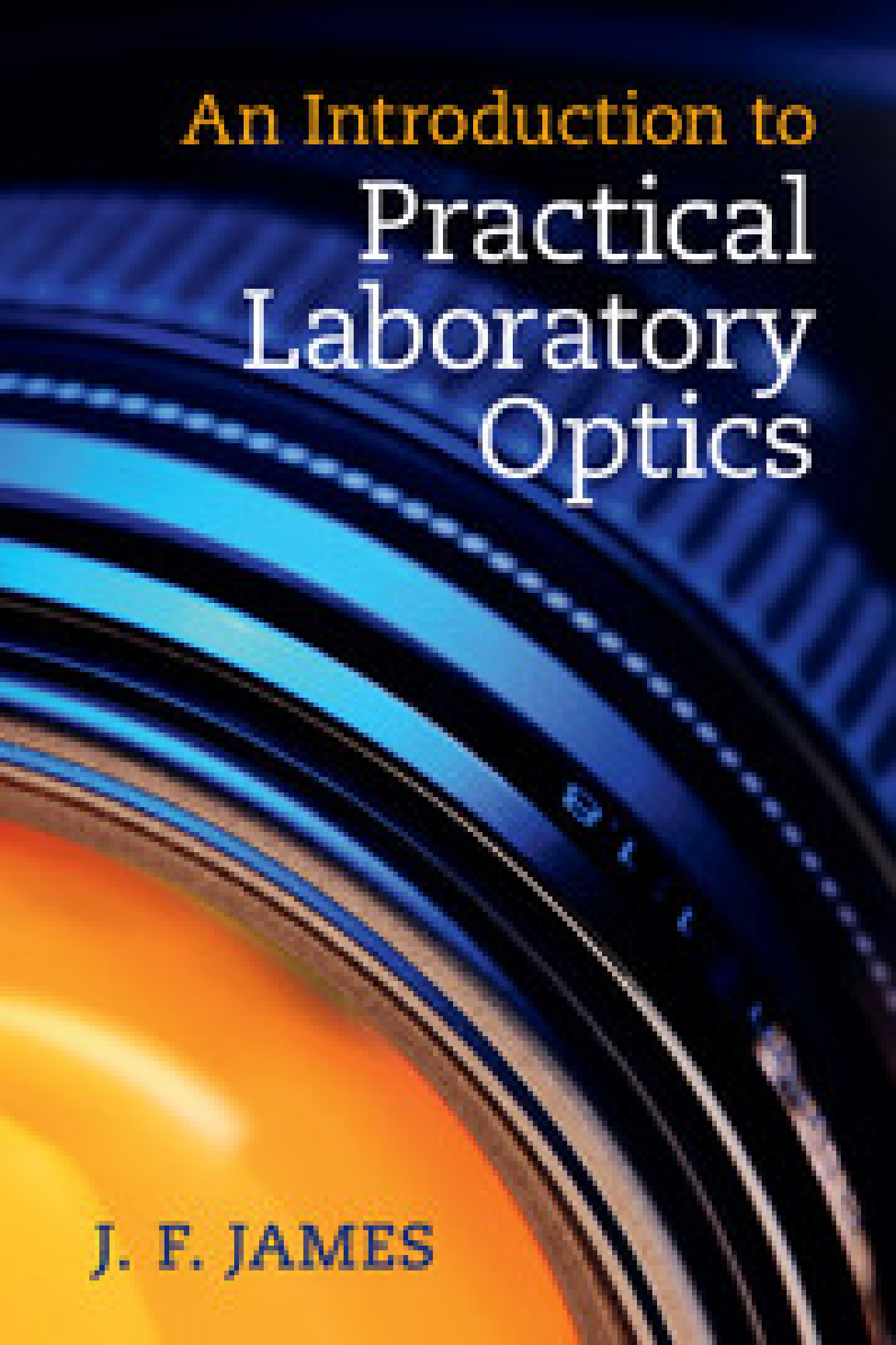
An Introduction to
Practical
Laboratory
Optics
J. F. JAMES

An Introduction to Practical Laboratory Optics

Aimed at laboratory workers and students taking laboratory courses in experimental optics, this book introduces readers to optical instruments and their uses.

The book explains the basic operation of lenses, mirrors, telescopes in the laboratory and under field conditions, how to use optical instruments to their maximum potential and how to keep them in working order. It gives an account of the laws of geometrical optics which govern the design, layout and working of optical instruments. The book describes the interactions of polarized light with matter and the instruments and devices derived from this, and discusses the choice of spectrometers and detectors for various spectral regions, with particular attention to CCD cameras.

The emphasis throughout is on description, with mathematical precision confined to the appendices, which explain the ray transfer matrix and outline the Seidel theory of optical aberrations. The appendices also introduce Fourier methods in optics and Fourier transform infra-red spectrometry.

J. F. JAMES has held teaching positions at The Queen's University, Belfast, and the University of Manchester, and is one of the pioneers of Fourier spectroscopy. He is the author of *A Student's Guide to Fourier Transforms*, now in its third edition, and *Spectrograph Design Fundamentals* (Cambridge University Press, 2011 and 2007 respectively).

An Introduction to

Practical Laboratory Optics

J. F. JAMES

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107687936

© J. F. James 2014

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2014

Printed in the United Kingdom by Clays, St Ives plc

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data

James, J. F. (John Francis), author.
An introduction to practical laboratory optics / J.F. James.
pages cm
Includes bibliographical references.
ISBN 978-1-107-05054-9 (Hardback) – ISBN 978-1-107-68793-6 (Paperback)
1. Optical instruments–Handbooks, manuals, etc. I. Title.
QC371.J36 2014
681′.4–dc23 2014014037

ISBN 978-1-107-05054-9 Hardback
ISBN 978-1-107-68793-6 Paperback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Contents

	Preface	*page* ix
1	**Introduction: centred optical systems**	1
1.1	The common properties of optical instruments	1
1.2	Optical elements	1
1.3	Concepts in optical instrument design	3
1.4	Gaussian optics	11
1.5	Aberrations – a conspectus	13
2	**Telescopes and binoculars**	15
2.1	The visual telescope	15
2.2	Galilean telescopes	16
2.3	Keplerian refractors	17
2.4	Erecting telescopes	17
2.5	Binoculars	19
2.6	Applications of the refracting telescope	21
2.7	Reflecting telescopes	22
2.8	The Newtonian telescope	23
2.9	The Cassegrain telescope	23
2.10	Refractors versus reflectors	26
2.11	Focal reducers	27
2.12	Image rotators	29
3	**Eyepieces, eyes and colour**	30
3.1	Eyepieces	30
3.2	Eyepiece properties	32
3.3	Astronomical photography through an eyepiece	35
3.4	Visual observation	36

3.5	The eye as an optical instrument	38
3.6	The perception of colour	39
4	**Cameras and camera lenses**	43
4.1	Introduction: the photographic lens	43
4.2	The double Gauss lens	44
4.3	Telephoto lenses	45
4.4	Wide-angle lenses	45
4.5	Fish-eye lenses	47
4.6	Zoom lenses	49
4.7	Practical aspects	49
4.8	Architectural photography	51
4.9	Schlieren photography	51
4.10	The optical transfer functon	54
4.11	The Boys points of a lens	55
5	**The scientific CCD camera**	57
5.1	The traditional silver halide camera	57
5.2	Scientific digital photography	60
5.3	Calibration	62
5.4	Radioactive lenses	62
6	**Spectrometry**	64
6.1	The prism spectroscope	64
6.2	The direct vision spectroscope	69
6.3	Diffraction grating spectrometry	70
6.4	Grating mountings	71
6.5	Concave gratings	76
6.6	Illumination of a spectrometer	77
6.7	Grating ghosts	80
6.8	The alignment of a grating spectrograph	81
7	**Interferometers and their uses**	84
7.1	Interferometry	84
7.2	Oscillation, phase and phase-difference	84
7.3	Coherence	85
7.4	Division of wavefront	86
7.5	Division of amplitude	88
7.6	Interferometers	88
7.7	The Fizeau interferometer	90
7.8	The Michelson interferometer	91
7.9	The Fabry–Pérot étalon	93

7.10 The Mach–Zehnder interferometer 99
7.11 The Sagnac interferometer 101
7.12 Fine displacement measurement 102
7.13 Optical resonators 103

8 Electro-optical effects and their practical uses 104
8.1 Introduction: polarized light 104
8.2 Polarization effects 105
8.3 Polarizing elements 108
8.4 Laboratory uses of polarized light 109
8.5 Faraday effect 110
8.6 Kerr effect 111
8.7 Pockels effect 113
8.8 A practical example of applied polarizing effects 113

9 Microscopes and projectors 116
9.1 Projectors 116
9.2 The Fresnel lens 118
9.3 Microscopes 119
9.4 Köhler illumination 120
9.5 Travelling microscopes 122

10 Siderostats and coelostats 124
10.1 The celestial sphere 124
10.2 Equatorial telescope mountings 124
10.3 Coelostats 126

11 The detection and measurement of radiation 130
11.1 The geometry of radiation measurement 130
11.2 The detection of radiation 132
11.3 The classification of radiation detectors 133
11.4 The far infra-red 133
11.5 The middle infra-red 134
11.6 The UVOIR region 134
11.7 The far ultra-violet 136
11.8 Types of spectra 137
11.9 Black body radiation 138

12 Practicalities 141
12.1 Cleaning optical surfaces 141
12.2 Mounting optical parts 144
12.3 Optical filters 145

Appendix A **Gaussian optics** 149

Appendix B **Optical aberrations** 162

Appendix C **A brief introduction to Fourier optics** 174

Further reading 181

Index 183

Preface

The most famous and illustrious mis-user of an optical instrument is Sherlock Holmes. There is an iconic figure of him with Inverness cape, deerstalker hat and calabash pipe, peering though a magnifying glass, the latter held at arm's length, to inspect a possible blood stain.

This is the wrong way to use a magnifying glass – which incidentally should always have a plano-convex lens rather than the biconvex lens with which many cheap versions are provided. The glass should be held close to the eye, plane side facing, and the object brought in until it is in clear focus at a comfortable distance for viewing. This gives the clearest image, the widest field and the minimum of optical aberrations. It is the attention to small detail like this which helps ensure success. Watchmakers do it properly, with a *loupe*, a lens held, like a monocle, in the eye socket.

In experimental science, especially in physics laboratories, it is sometimes found, when beginning a new piece of basic research, that no appropriate apparatus exists and that it is necessary to improvise. The traditional laboratory stand-and-clamp then comes into its own, followed, after some experimentation, by a properly designed system with an optical bench or table with lenses, mirrors and other basic optical elements for measuring and analysing radiation. Skills in the design and assembly of such provisional devices are part of the true experimenter's art.

In any field of science, a large part of experimentation involves light, be it photography, photometry, polarimetry, microscopy or spectroscopy, and the proper selection and application of optical elements and devices may have a great effect on the satisfactory outcome of experimental work. Much can be done with standard, versatile 'common-user' instruments such as cameras and microscopes, but all too often there is a particular measurement to be made where apparatus must be designed and constructed before the experiment can be done.

No mention is made here of individual manufacturers of optical equipment. Even famous firms come and go or are swallowed by other, bigger fish. But they advertise in the various scientific journals and their catalogues provide valuable information about their products and much useful background information on optics generally.

What follows here is an account of the basic principles of common optical instruments, the elementary laws of optics affecting them and descriptions of some of the ingenious devices which have been invented over the last few centuries and which are of particular use in the laboratory, in the workshop and in the field. [1]

I make no apology for presenting theorems and useful equations without proof. This is a handbook, not a treatise, and there are many learned works on library shelves which will supply the deficiency if a more thorough explanation is needed. The list of further reading at the end of this book will point the inquirer in the right direction.

Finally, a warning! Geometrical optics is a dangerous subject and I have left it to the end. For most sensible people it is a miserably boring agglomeration of tedious, elementary calculation, arithmetic and geometry, but to others it is an obsession like chess or golf or computer programming. The victim is sucked into a quagmire of invention, ingenuity and improvement, forever trying to achieve the perfect lens or telescope or whatever. The lucky designer is the one who knows when to stop calculating and start building. Yet there is an immense sense of achievement when the last element of the new system is put in its place, the last adjustment is made, and one's final fears evaporate as a clear, sharp image appears in its proper place, just as calculated. This is invention at its most satisfying.

Beppu, Japan
Beppu, Japan, October 2013 J. F. James

[1] 'Field', here, is the usual euphemism for mosquito-ridden swamp or bare, windswept mountain-top where some deranged scientists choose to do their data-gathering.

1

Introduction: centred optical systems

1.1 The common properties of optical instruments

Optical instruments come in all shapes and sizes, from fly-on-the-wall surveillance cameras to 10 metre segmented astronomical reflecting telescopes, and in shapes from microscopes to sextants to periscopes to spectrographs to cine-projectors.

Whatever their purpose, they all have two things in common.

(1) They are *image-forming* devices, intended to make a picture, to form an image of a luminous source. The image may be on a cinema screen, on a photographic emulsion, on a CCD surface or on the retina of an eye.
(2) They are, with one important exception, *centred systems*. That is to say they comprise a series of curved surfaces of transparent materials or reflecting materials or both. The centres of curvature of the various elements all lie on a straight line called the *optic axis*. Light passes from the object, through successive elements until it emerges to form an image.

This is a slight over-simplification of course. There are occasional plane reflectors along the path, as in a periscope for example, but these are for convenience rather than for any peculiar optical properties they possess.

1.2 Optical elements

There are four basic optical elements: the lens, the mirror, the diffraction grating and the prism. What follows now concerns the first two of these. The others have chapters of their own.

1.2.1 The lens

Lenses have been known since antiquity. The word lens is the Roman word for a lentil. A lens may be simple as in a magnifying glass or composed of several *elements* as in a telescope or camera, and the word is used indiscriminately. We are concerned for the moment with the simple element, which is a disc of glass with two polished surfaces, one or both of which may be curved. The curved surfaces are usually spherical, although non-spherical surfaces are becoming commonplace at the time of writing.

Its chief properties are (a) its *aperture* and (b) its focal length. The aperture is the diameter of the transparent area. The focal length is the distance from the *vertex* [1] of the lens of the image formed of an object at $-\infty$. If the curvature is concave the focal length is negative and no real image can be formed. Negative lenses are almost invariably part of a more complex lens system which does form a real image.

Lenses may be described as *biconvex*, *plano-convex* or *meniscus*, depending on the direction of curvature of the two surfaces. Two lenses may have the same aperture and focal length, but focal length depends only on the *difference* in the two curvatures, so that two lenses of the same focal length but different front surface curvature may have different image-forming properties. As a simple guide to these image-forming properties, remember that as a general rule refraction should be equally distributed among the various surfaces for a good sharp image. For example, the image quality of a plano-convex lens will depend markedly on which surface faces the object side.

Lenses are traditionally made of glass and much research has gone into making high quality optical glass, free from bubbles, striae [2] and stones (which cause scattering), free from stresses (which cause birefringence) and of high homogeneity (to avoid aberrations). Rare earths such as thorium and lanthanum have been incorporated to achieve high index and low dispersion (both desirable properties), and refractive indices between 1.4 and 1.9 are available (at a price); but polymer materials such as polystyrene and polycarbonates, with refractive indices $\sim$1.6, are also available for ordinary optical instruments.

Instruments which consist only of lenses, such as refracting telescopes and microscopes, are called 'dioptric' systems; those comprising only mirrors are 'catoptric' and when both are used they are 'catadioptric'. Reflecting telescopes and Cassegrain long-focus camera lenses come into this latter category.

[1] The word is used loosely here. Strictly *surfaces*, not lenses, have vertices, and the vertex is the point where the optic axis meets the surface.

[2] Known as sleaks in the trade.

1.2.2 The mirror

Early mirrors were of bronze. In the seventeenth century alloys were found which were hard enough to be ground and polished like glass, and eventually *speculum metal* was perfected, an alloy of approximately two-thirds copper and one-third tin. With a high polish it reflected about 70% of the incident light. In 1835 Justus von Liebig (1801–1873) of the University of Giessen described a method for precipitating colloidal silver from an ammoniacal solution of silver nitrate [3] and it then became possible to deposit highly reflecting silver surfaces on to polished glass. These surfaces, when newly deposited, had a reflectivity of about 92% but as the industrial revolution expanded they tended to tarnish in the sulphurous atmosphere which developed around coal-burning cities.

With the development of vacuum techniques, evaporation of metals became possible and reflecting surfaces of aluminium were found to be chemically durable and with a reflectivity of about 88%. These were improved further by *overcoating* them with evaporated films of silicon monoxide to give a hard and washable surface finish. Reflectivity improves in the infra-red, and gold surfaces, as well as being chemically durable, show a reflectivity in the region of 99%. In the far ultra-violet aluminium may be overcoated with magnesium fluoride to improve its reflectivity which otherwise falls rapidly as the wavelength shortens.

Mirror surfaces may be plane, spherical, paraboloidal or elliptical, and some current mirror-forming techniques allow various, even more elaborate surfaces to be formed by computer-controlled grinding and polishing, chiefly for large astronomical telecopes.

1.3 Concepts in optical instrument design

Let us consider first some features common to all optical instruments, which we need in order to understand how they work. We begin with the description of light.

1.3.1 Rays of light

First of all is the assumption that light travels in straight lines, *rays*, emitted by an object, passing through various lenses and mirrors and converging finally to form an image on a focal surface: a screen, a photographic plate, a CCD or a retina.

[3] An early version of Tollens' reaction.

This assumption is one of the 'convenient fictions' of physics, useful for describing what happens and for predicting the behaviour of light in instrument design; but the model fails in the end to take account of diffraction and interference in the *minutiae* of image formation. Modern optical design requires the *wave theory* of light propagation to account for the fine detail in an image.

Wave theory supposes that light is an electromagnetic field which propagates outwards from a point source in *wavefronts*, spherical surfaces on which the electric field is a maximum, and which expand outwards at the speed of light. The perpendicular distance between successive wavefronts is the wavelength, measured either in angstrom units (1 Å $= 10^{-10}$ m, roughly the diameter of an atom) or in nanometres (10^{-9} m), the standard SI unit. [4]

A lens or a concave mirror can be used to convert part of these diverging wavefronts into spherically converging wavefronts. In a lens, the process is governed by *refraction*, the speed of light being less in a dense transparent medium than in air.

However, it is easier to describe the process of refraction by connecting an object and its image by a ray, the direction changing at each refracting surface.

1.3.2 Optical ray diagrams

It is a convention, not always observed, that diagrams showing rays passing through optical systems show the optic axis horizontal, and show rays coming from the left, the object side or input side, and leaving the system on the right, the image or output side. The usual Cartesian formalities are observed and lengths measured to the left of the lens – the input side – are considered to be negative and those measured to the right are considered positive.

1.3.3 Refraction and the law of sines

The earliest known exposition of the law governing the refraction of light at a surface was by the mathematician Ibn Sahl (940–1000) of Baghdad in the tenth century. It was rediscovered several times in Europe following the Renaissance and today is known in English-speaking countries as *Snell's law* after Willebrord Snel van Royen (1580–1626) of Leiden. It is also more simply known as the 'law of sines'.

In 1657 Pierre de Fermat (1601–1665) published his *Principle of Least Time*, which asserts that light travels from object to image by that path which takes

[4] Physicists tend to use nanometres, while astronomers cleave to the old angstrom unit. Both units will be found in this book.

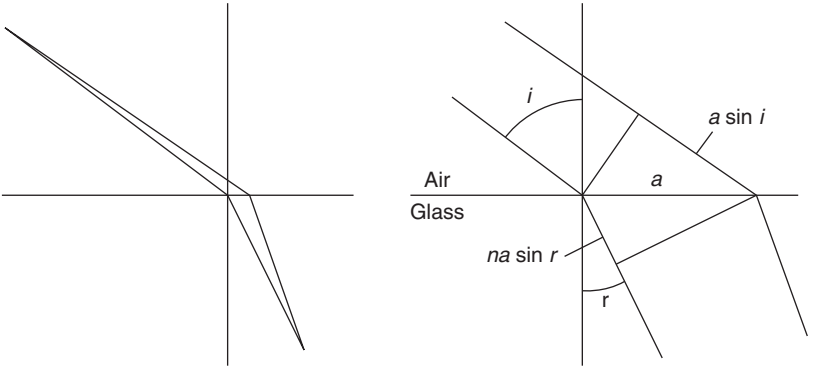

Figure 1.1 The law of sines (Snell's law). When the optical path length is near a minimum, a small change in direction makes no change in the total path length. The small extra path in air is exactly matched by the small diminished path in glass. When the total change is zero, the law of sines follows immediately.

the shortest time. From Fermat's principle it is a simple exercise in differential calculus [5] to derive Snell's law, that the sine of the angle of incidence divided by the sine of the angle of refraction is a constant, the *refractive index*, usually [6] denoted by the letter n:

$$n = \frac{\sin i}{\sin r}$$

The derivation of the 'law of sines' is simple. The small change of *time*, Δt, is given by:

$$\Delta t = a \sin i - n.a \sin r$$

since the speed of light is less in glass by a factor n. At a minimum the change in time is zero, which gives the 'law of sines' immediately.

1.3.4 Optical paths

We occasionally need three different path lengths to describe the passage of light through refracting media.

(1) The *geometrical path*, the path actually laid out on the drawing board.
(2) The *optical path*, the geometrical path multiplied by the refractive index of the medium through which each section of the path the ray is travelling.

[5] And is evidence that Fermat had the idea of the *differential* calculus before Newton and Leibnitz, even though he never made the connection with integration.
[6] But in many older books by the Greek letter μ.

Alternatively (and more accurately), it is the number of wavelengths of light between object and image, multiplied by the vacuum wavelength of the light. All the rays between an object point and its image point must have the same optical path length.

(3) The *reduced path*, the geometrical path *divided* by the refractive index. This is the path length you would measure if you were using a rangefinder to find the apparent distance to an object at the bottom of a swimming pool.

1.3.5 Apertures and stops

The *aperture* of a lens is self-evident. It is the diameter of the transparent part of the lens. The *physical* diameter is a millimetre or two greater than this to allow for 'turn-down' at the circumference in the manufacturing process, and to allow for the 'cell' or retaining device which holds the lens in place.

Stops are opaque screens normal to the optic axis, with circular holes to limit the size of the ray bundle passing through the instrument. They may fulfil several purposes. They may be positioned to control the aberrations of the system, they may be 'anti-glare' stops, used to reduce scattered light from various internal surfaces which would otherwise reduce the contrast of the image and prevent accurate photometry, and if they are variable in diameter as in a camera iris diaphragm, they control the intensity of light forming the final image.

1.3.6 Pupils

There is always one defining stop in an optical instrument, the *iris*. It may simply be the rim of a lens or it may be somewhere inside the sytem, in which case its image, seen through all the elements on the input side, is the *entry pupil* of the system and the corresponding image on the output side is the *exit pupil*. Pupils are not necessarily inside the system. In a telescope for instance, the exit pupil is several millimetres to the right of the last element of the eyepiece: it is where you put your eye-pupil when inspecting a distant object. All the rays of a ray bundle must pass through the pupils of a system.

1.3.7 Relative aperture

This is popularly referred to as 'focal ratio' or 'F-number' and is denoted by F. It is the ratio of the lens focal length to its aperture. As such it is a measure of the intensity of light at an image point and consequently is useful in photography and photometry. It is usually expressed as a fraction of the focal

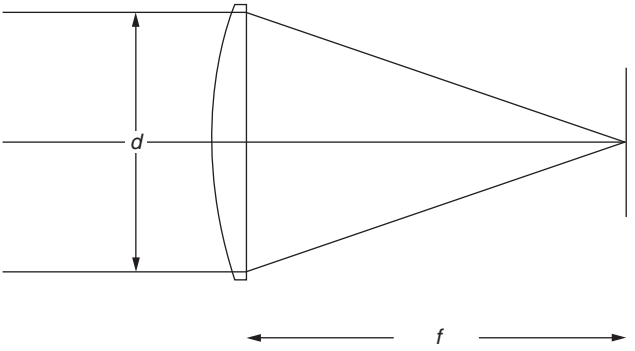

Figure 1.2 The focal ratio or *relative aperture* or F-number. This is the ratio of the focal length of the lens to its clear aperture. In general it is defined as the ratio of the focal length to the pupil diameter and is expressed as, for example, f/8 or f/6.3.

length so that, for example, a lens with relative aperture of f/6 has a clear aperture of one-sixth of its focal length. In complex lenses with many elements it is the exit *pupil* diameter as a fraction of the focal length which measures the focal ratio.

1.3.8 Numerical aperture

In a microscope the object is extremely tiny. In order to see a great enlargement, the light from the object must be collected from the greatest possible solid angle and turned into a large image at a much greater focal ratio. For microscope objective lenses we refer to its *numerical aperture*, defined as:

$$\mathrm{NA} = n \sin \theta_m$$

where θ_m is the angle which the ray from the object to the edge ('the margin') of the objective lens makes with the optic axis *at the object*, and n as usual is the refractive index in the object space, i.e. *before* the first surface of the lens. Numerical aperture is considered at greater length in the chapter on microscopes. So-called *immersion objectives* in microscopes take advantage of the refractive index of oil to increase the NA by a factor n. For most practical purposes, the numerical aperture is related to the relative aperture or F-number by

$$\mathrm{NA} = 1/2F$$

1.3.9 Étendue

This indicates the radiated power which can be transmitted through the system from a distant extended source of light. It refers mostly to photometers and radiation detectors and is a measure of the flux-collecting capability of an optical instrument. It can be defined in two ways.

(1) It is the exit pupil area multiplied by the solid angle subtended there by a detector.
(2) It is the sensitive area of a detector multiplied by the solid angle subtended there by the exit pupil.

One must be careful here: each pixel of a CCD is a separate detector. The rate at which information is gathered in a properly designed experiment will depend on the *total* étendue of the device, i.e. on the total sensitive area of the CCD multiplied by the solid angle subtended by the exit pupil.

The concept of étendue as the product of area $\times$ solid angle is one of the most profound and important concepts in optics and radiometry, and one which determines the design of many types of optical instrument. Étendue is an invariant of an optical system, and is *conserved* throughout various stages of re-imaging.

1.3.10 The Helmholtz–Lagrange invariant

This is best described with a diagram (see Figure 1.3). If objects are imaged and then re-imaged throughout a system, then the height H_i of the ith image, multiplied by the angle from the image base to the rim of the next aperture A_{i+1}, and multiplied by the refractive index of the medium between, is a constant throughout the system. There are in fact two such invariants in any instrument, defined by two orthogonal planes containing the optic axis. If there is

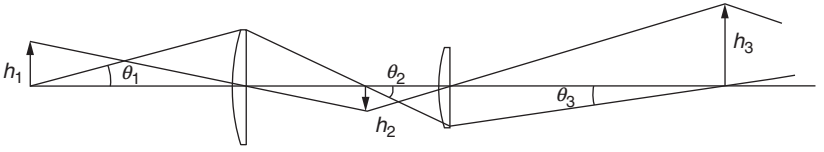

Figure 1.3 The Helmholtz–Lagrange invariants of a centred system. The product of the image size, h_n and the angle θ_n which the next, or previous, aperture subtends at the image position, is a constant of the system. There are two of these invariants for any system, but if there is circular symmetry about the optic axis, they are identical. In an asymmetrical system such as a spectrograph, they are not.

rotational symmetry they are the same but if the instrument is *anamorphic* they will be different, each being separately conserved though the various stages. Taken together they represent the étendue.

A simplified and practical version is to say that the image size divided by the focal ratio is a constant. People who photograph through telescopes will understand this one: if you want a brighter image, i.e. a smaller focal ratio, you have to accept a smaller image. The amount of light coming through is always the same. Étendue is conserved.

1.3.11 The vignette

This is a term which derives from the word 'vine', which was used classically, suitably entwined, as a tasteful decorative border to a fresco, and subsequently adopted by Victorian photographers to describe the border of a portrait which softened the hard outline by allowing the image to fade gradually at the edges. In optics a vignette is something highly undesirable, and usually indicates that an instrument has not been designed properly or is being used beyond its design limits.

An optical system is said to vignette when there is no single controlling stop to regulate the passage of light through it, so that the intensity of an image varies from point to point from the centre outwards.

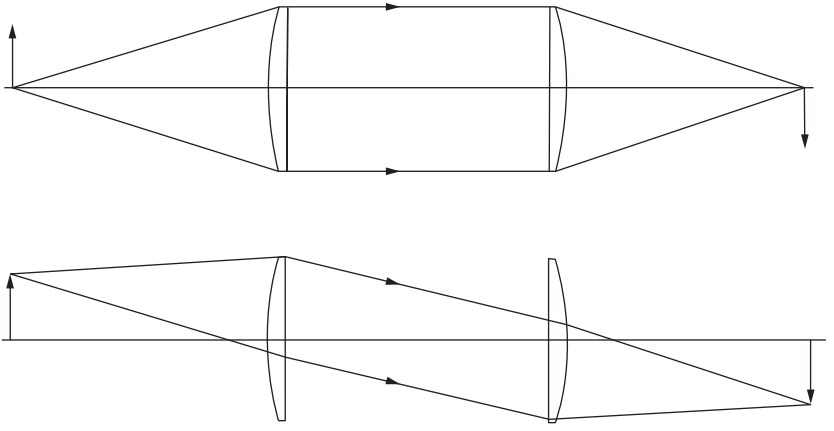

Figure 1.4 The vignette. The flux of light through this system decreases as the obliquity of the ray bundle increases so that the second lens cannot be completely filled with light. A stop half-way between the two lenses would ensure that the intensity is constant over a wide field, and a field stop may be placed on the image plane to mark the limit of the field of uniform illumination.

1.3.12 Conjugate points

These are defined as points connected by ray bundles. A point on an object is *conjugate* to the corresponding point on its image. In this sense an extended image is, in the mathematical sense, a *map* of the object because there is a one-to-one relation between points of each. The idea is not constrained to object and image points. Points on the entry and exit pupils of a system are conjugate. Light which has passed one point must also necessarily pass through its conjugate. When an instrument has several optical stages it is important that the exit pupil of one be the entry pupil of the next to ensure that all light passes through the system without vignetting.

1.3.13 Ray bundles

Rays from an object point to its conjugate image point spread out and pass through different parts of the optical system. All have the same optical path length [7] and collectively form a ray bundle. An important member of this bundle is the *chief-ray*, generally the central ray of the bundle, which is distinguished by the property that *the chief-ray of a bundle crosses the optic axis at the pupils of the system*.

The chief-ray is of particular importance when interference filters are used to do spectrometry. These filters have a narrow spectral pass-band and the peak wavelength transmitted depends on the angle of incidence. The pass-band of the filter moves to shorter wavelength as the angle of incidence increases. The effect is serious, so that it is important that the chief-ray of every ray bundle be everywhere normal to the surface as it passes through the filter.

1.3.14 Telecentricity

A system is said to be *telecentric* when all the chief-rays entering or leaving the system are parallel to the optic axis. It can be ensured for rays entering the sytem by putting a stop at the rear focus of the objective lens. It is of particular practical importance in instruments such as profile projectors to ensure that a true silhouette of an object is produced at the image position. Such systems are said to be 'telecentric on the object side'. In Figure 1.5 it is necessary for the chief-rays of all bundles to pass through the filter on the right parallel to the optic axis. It is achieved here by placing a lens near the focal plane, of the same focal length as the objective on the left.

[7] Give or take a few wavelengths, owing to aberrations.

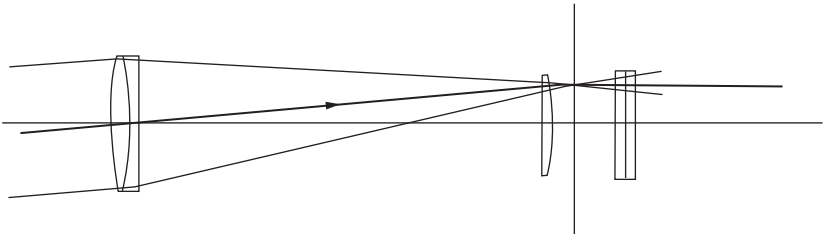

Figure 1.5 The chief-ray of a bundle. The extra lens in this simple system ensures that all ray bundles passing through the interference filter on the right have their chief-rays parallel to the optic axis. This is always necessary when an interference filter is used to restrict the photograph to a small band of wavelengths. This system is said to be *telecentric on the image side*.

1.3.15 Collimated light

A collimated beam of light is one in which all ray bundles have come from points on a luminous object at $-\infty$. A *collimator* is a lens with a point source or possibly a slit at one focus, so that light coming from the source or through the slit may be said to be collimated.

1.4 Gaussian optics

There are simple equations describing the passage of rays through lenses and mirrors from object point to its conjugate image point, but they are valid only if the angles involved are vanishingly small and if both apertures and fields are infinitesimal. The rules are that $\sin\theta \sim \theta$ and $\cos\theta \sim 1$. This is not the place to treat them in detail, and the equations are given in full in Appendix A. The consequences of these rules are of practical value and help to make a preliminary layout of a new instrument.

As usual the distance from a lens to an object point (to its left) is u and to its conjugate image point v. If the object is at $-\infty$ the image is at a distance f from the lens and f is defined to be the focal length of the lens. Likewise, if the object is at a distance $-f$ to the left of the lens, the image is at ∞. The two focal lengths are the same. But be careful! This is only true if both object and image are in a medium of the same refractive index, such as air. If one or the other is in water for example, these remarks are no longer true. The two focal lengths will be different.

1.4.1 Cardinal points and planes

In Gaussian optics any image-forming system, whether lens, telescope or projector, however complex, can be represented by four planes all perpendicular

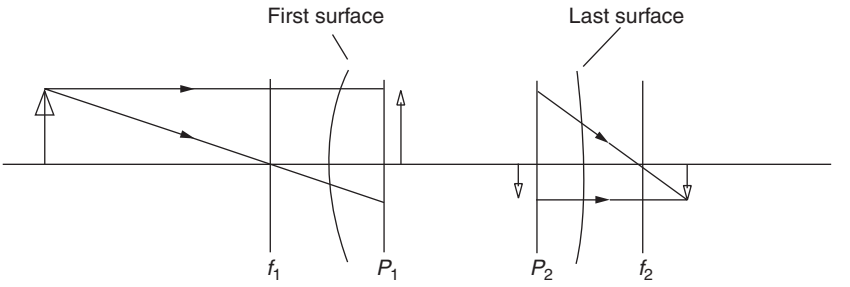

Figure 1.6 The cardinal points and planes of a centred system. For purposes of ray-tracing, the system is fully defined by four planes, P_1, P_2, f_1 and f_2. The actual *geometrical* positions of the first and last surfaces are immaterial.

to the optic axis. These are the two *principal planes* and two *focal planes*. Where these planes intersect the axis are the four *cardinal points* of the system, which are labelled[8] f_1, f_2, P_1 and P_2. They may occur in any order and their positions depend on the particular arrangement of the various elements of the system. They may even be at infinity. They are important because with them the rules for constructing ray paths are simple and sufficient for a preliminary drawing of the optical layout of the instrument. The rules are the following.

(1) A ray entering from the left parallel to the optic axis arrives at the first principal plane P_1 at height h. It leaves the system at the second principal plane P_2 at the same height h and passes through the second focal point f_2.
(2) A ray entering from the left and passing through the first focal point, f_1, arrives at the first principal plane P_1 at a height h' and leaves the system from the second principal plane P_2 at the same height h' and parallel to the optic axis.

These two rules are sufficient *in the Gaussian approximation* for the construction of the image point of a given object point.

There is a third pair of cardinal points, the *nodal points* of the system, defined by the rule that a ray arriving at point N_1 leaves the system at point N_2 at the same angle to the optic axis. For most practical purposes the nodal points coincide with the principal points P_1 and P_2, but again this is not so if the refractive indices in object-space and image-space are different. A useful practical property of the nodal point is that if the whole instrument is rotated about an axis passing through the nodal pont and lying in the nodal plane, the image will not move.

[8] They are sometimes labelled as P, f, P' and f'.

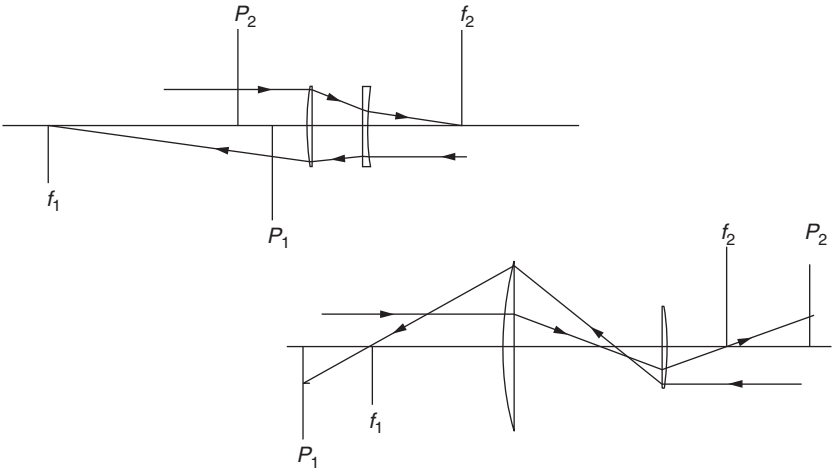

Figure 1.7 Two examples of centred systems showing the positions of their cardinal points. The rules for drawing rays are always the same but the *order* of the four cardinal points may vary.

Again in Gaussian optics the principal points of a thin lens are assumed to be coincident.

1.5 Aberrations – a conspectus

In a perfect image-forming instrument every point of an object would be conjugate with a corresponding point in the image. All the ray paths between the two would have exactly the same optical path length. This does not happen, because geometry and the law of refraction conspire to make the behaviour of rays highly complex, and only when the angles of incidence and refraction are vanishingly small – the Gaussian limit in other words – is there good approximation to the ideal. In practice it is necessary to consider finite apertures and finite field angles, and the approximation now is that $\sin\theta = \theta - \theta^3/6$, $\cos\theta = 1 - \theta^2/2$ and $\tan\theta = \theta + \theta^3/3$. This is sometimes called 'third-order theory' or 'the Seidel limit'.

The subject of optical aberration was codified in the nineteenth century by Ludwig von Seidel (1821–1895) of the University of Munich, who treated aberrations in eight categories: five *geometrical aberrations*, two *chromatic aberrations* and *diffraction*. They are considered in detail in Appendix B, where there are formulae for computing their severity. The severest is *spherical aberration*, a consequence of using spherical glass surfaces to do the

focusing, in which rays passing through the edge of a lens do not arrive at the same focal point as rays close to the optic axis. There are various tricks for correcting this. Next is *coma*, an off-axis aberration, where again the marginal rays arrive at a different place from the paraxial rays, giving a comet-like 'tail' to the image of a distant point source.

It is a legend among optical designers, that 'if you can correct the sph. ab. and the coma the other aberrations tend to disappear of their own accord'. It is only slightly true.

Astigmatism, like coma, is an off-axis aberration. It is the change of focus for two-dimensional pencils of rays, one pencil in the plane of the optic axis and the oblique axis, the other perpendicular to the first. The result is that ray pencils either focus as a short line reaching radially outwards from the axis (the 'sagittal focus') or a short line perpendicular to this (the 'tangential focus'). A wheel with radial spokes will focus the circumference at the tangential focus and the spokes at the sagittal focus.

The last two geometric aberrations, field curvature and distortion, are symmetrical aberrations. The first simply means that the focal surface is not a plane but a bowl – the retina of the eye is the classic example. Distortion means that a square is imaged either as a barrel or as a pin-cushion. Distortion happens when the focal length of the lens varies with the off-axis angle, increasing for pin-cushion distortions and decreasing for barrel distortions. 'Symmetrical' lenses, so-called because they comprise two sets of elements equally disposed mirror-image fashion before and after the controlling iris diaphragm, tend to be free from distortion.

The two chromatic aberrations are longitudinal, where the focal length varies with wavelength, and lateral, where although the focal plane is the same for all colours, the principal plane position is different, so that the focal length varies with wavelength and the image size changes. Both of these aberrations result in 'colour-fringing' where a dark image has a coloured surround.

All these aberrations are discussed in more detail in Appendix B.

2

Telescopes and binoculars

2.1 The visual telescope

This is an elementary type of optical instrument.

The historically acknowledged function of a telescope is to produce an enlarged virtual image of a distant object for inspection by the observer's eye. Both object and image are at $-\infty$ in normal focusing mode, but by shortening the tube a little the image can be brought nearer to the observer's eye to allow for myopia. In effect the objective forms a real image of a distant object and the eyepiece re-collimates the light so that the final image, like the object, is at $-\infty$. The lens of the observer's eye then focuses this virtual image on to the retina.

An unusual but equally valid alternative is to regard a telescope as a device which puts an enlarged, real image of the observer's eye-pupil at the telescope's objective. If the telescope is properly designed, the image of the eye-pupil will exactly correspond in position and diameter to the clear aperture of the objective. The enlarged, virtual eye would then see the distant object exactly as the actual observer sees it, but in appropriately greater detail.

What all this implies is that the observer's eye-pupil and the telescope objective are conjugate points in the complete optical system of telescope and eye.

The same may be said of a periscope: it puts an image, not necessarily enlarged, of the submarine commander's eye at the upper end of the periscope tube, clear of the water. Instead of providing high magnification, it solves the problem of how to pass an image, with a wide field of view, down a long narrow tube. It does so by re-imaging the field and re-re-imaging it a sufficient number of times until an image of the entry pupil (the exit pupil in other words) appears at the other end. One way to do this is to put a series of telescopes back-to-back all the way down the tube, the exit pupil of one coinciding with the entry pupil of the next.

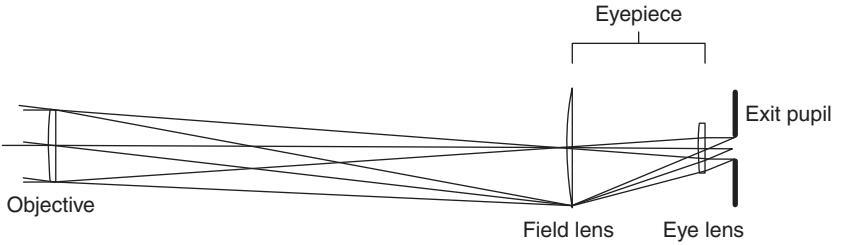

Figure 2.1 A simple refracting telescope. The objective produces a real image near the field lens of an object at $-\infty$. The field lens ensures that all ray bundles forming the image are diverted to pass through the eye lens. The eye lens re-collimates the light so that the image appears to the observer to be at $-\infty$. The field lens and eye lens together form an image of the objective lens beyond the eye lens. This is the *exit pupil* of the system, the place where the observer's eye-pupil should go in order that the eye be filled with light. The distance from the last surface of the eye lens to this image is called the *eye-relief*.

The magnification of a telescope is a ratio of angles. The Moon for example subtends an angle of about 31 minutes of arc [1] more or less, at an observer's unaided eye. With a $\times 7$ binoculars the image will subtend about 3.5°. The magnification is the ratio of the focal lengths of the objective and the compound eyepiece. It is also the ratio – and this is easier to measure – of the entry and exit pupil diameters. Since there is always eye-relief, a ruler applied across the exit pupil will give its diameter and this can be compared with the objective diameter to give the magnification.

Thus a 7×50 mm binocular telescope has an exit pupil diameter of about 7 mm, corresponding to the pupil diameter of the dark-adapted eye. This is slightly more than is necessary as very little light passes through the outer half-millimetre of the eye-pupil and a a 7×42 mm telescope would work just as well. The extra half-millimetre is to allow for the trembling of the eye in its socket.

2.2 Galilean telescopes

The telescope first appeared early in the seventeenth century, reputedly the invention of Hans Lippershey of Delft, but was put to good astronomical use by Galileo Galilei (1564–1642) of the University of Padua. Its objective was a

[1] The angle is variable, not only because of the excentricity of the Moon's orbit but because the Moon's apparent diameter, hunters' and harvesters' opinions notwithstanding, is greater when she is at the zenith than when she is near the horizon. This is known as the *augmentation of the Moon* and it happens because the observer is 4000 miles nearer to the Moon when she is overhead. The effect amounts to about 31 seconds of arc.

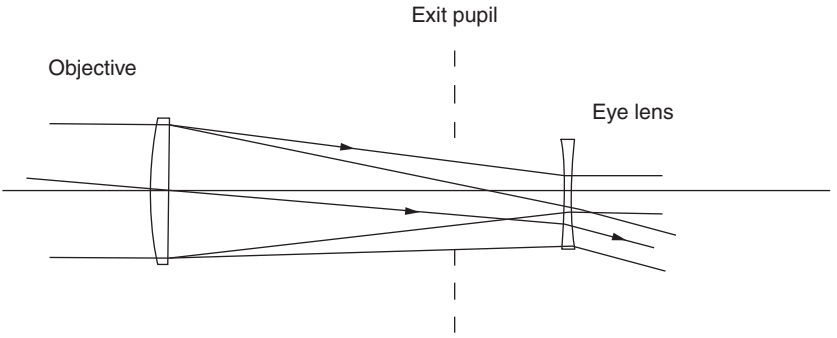

Figure 2.2 The Galilean telescope. This is the original form of telescope used by Galileo Galilei (1564–1642) to discover the moons of Jupiter (and the 'handles', as he put it, on Saturn!). Its exit pupil is inside the tube and inaccessible to the observer's eye-pupil. Consequently it is severely limited in its field of view and image quality, and is chiefly of historic interest.

convex lens, followed by a concave lens as eyepiece, and it was of very simple construction. The arrangement suffers the major inconvenience that the exit pupil is inside the barrel of the instrument. There is consequently a very small field of view and such telescopes are capable of only low ($\times 3$) magnification without severe vignetting. They are rarely found these days except in cheap (and virtually useless) so-called 'opera glasses'. The sole merit of the system is simplicity: it gives an erect virtual image at $-\infty$ without complication.

2.3 Keplerian refractors

The idea of using two convex lenses to produce a magnified but inverted image originated with Johannes Kepler (1571–1630), and the compound eyepiece followed with Christiaan Huygens (1629–1695), which greatly improved the field of view. The inverted image was no embarrassment for an astronomer but most inconvenient for a seaman or a bird-watcher. The modern refracting telescopes which descend from these early ideas use several different methods to make the image erect.

2.4 Erecting telescopes

The earliest method for erecting the image was to insert a pair of lenses between the primary image and the eyepiece. These lenses together re-imaged

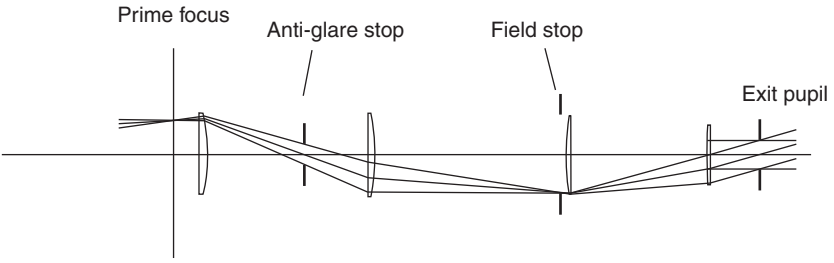

Figure 2.3 The erecting eyepiece. The figure shows the train of lenses which follow the prime focus. There is an extra pair of lenses to re-image the inverted image formed by the objective. It is a relay forming a new image usually but not necessarily at a magnification of -1, to be followed by an eyepiece (a Ramsden in the figure) in the usual way.

the primary image, generally (but not always) at magnification -1, ready for the eyepiece to re-collimate it for viewing. This method is still generally used for traditional extensible telescopes but never in binoculars.

2.4.1 Telescopic rifle sights

These are erecting telescopes of moderate magnification ($\times 4$ typically) distinguished by two things. There is a graticule at the prime focus which is moveable both vertically and horizontally in a transverse direction to make allowance for range and windage; and the eyepiece eye-relief is several centimetres to avoid eye damage at the recoil. The objective aperture is large enough to give a large exit pupil, both for ease of sighting and for operation in feeble light.

Rifle sights generally are of limited use except for their designed purpose.

2.4.2 Telescopic projection

An extensible telescope can be used to throw a real image of a distant object on to a screen. This is done by installing a white or translucent screen at a suitable distance for viewing (the choice of distance is at the disposal of the observer) and extending the tube until a distant scene is in focus on the screen. It is usual to exclude ambient light either by enclosing the rear end of the telescope in a box with a viewing aperture or by observing from a darkened room, and with a stand this is a suitable way of observing the Sun. With a moderately sized simple telescope an image 150 mm in diameter of such a bright object is readily available.

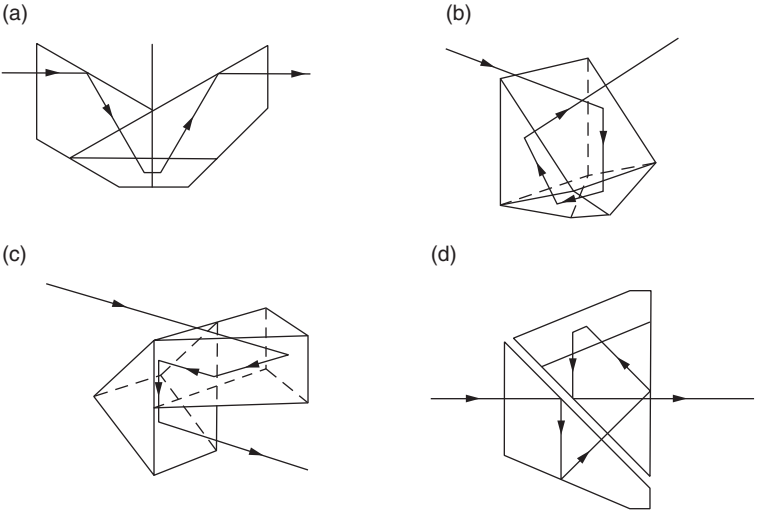

Figure 2.4 Image erecting prisms: (a) Abbé, (b) Amici, (c) Porro, (d) Schmidt–Pechan. These are alternative ways of producing an upright image in a telescope or binocular. They have the advantage of compactness, and shorten the length of the telescope tube substantially as well as saving weight. Notice that a roof-edge is included where necessary, to ensure that there is an even number of reflections. These prisms will *invert* but will not *rotate* an image.

2.5 Binoculars

These are almost as old as single Galilean telescopes, presumably because of the sound principle that two eyes are better than one. Rectification of the image is most frequently by *Porro prisms*, pairs of $45°-90°-45°$ glass prisms which reflect four times internally and displace the optic axis sideways. Alternatively some more expensive binoculars use inverting prisms, usually of the Schmidt–Pechan or Abbé–König type. The advantage of the latter is that the system, although more expensive to make, is fixed and rigid whereas the Porro prisms can be disturbed by a jolt and then need re-aligning by an expert.

The traditional sea-going binocular telescope has objectives with a 50 mm clear aperture and a magnification of $\times 7$. There are two reasons for this combination: firstly, this is the greatest magnification that can be comfortably used, hand-held, and any magnification greater than this results in a random image displacement too rapid for the eye to follow. There are modern exceptions of course, and image-stabilizing mechanisms allowing $\times 20$ are available – at a price. However, the second reason militates against this. For use at night a 7 mm exit pupil is needed to match the dark-adapted eye-pupil. The $\times 7$

magnification then requires objectives of 49 mm aperture. Consequently, a magnification of ×10 requires 70 mm objectives, and ×20 needs 140 mm objectives. The amount of glass in such lenses (not to mention the erecting prisms) makes the instrument too heavy to hold and a firm mechanical mounting with pan-and-tilt motion is needed.

2.5.1 Night glasses

How do they work? They do not make objects brighter. They make sure that your eye is filled with light – which is why the exit pupil is 7 mm diameter – and at the same time the magnification lengthens curves and straight lines, such as roofs or flag-poles, so that the mind is able to fill in the gaps in observation and imagine a straight line where there is no such image actually on the retina. To see the effect, take a pair of 7 × 50 binoculars on a really dark moonless night and look at a telegraph pole for example or a similar distant structure. There is more detail to be seen in the structure simply because the lines are longer.

2.5.2 Magnification and aperture

Binoculars are of course used for other purposes and in daylight (bird-watching, deer-stalking etc.) where a 4 mm exit pupil is sufficient. Again the ×7 magnification limit holds for hand-held use, but an 8 × 30 system is feasible and is commonly available. A 12 × 50 binocular will almost certainly need a fixed mounting – a tripod probably, although resting the instrument on a bean-bag has been used before this. For more convenient viewing, especially from a prone position, an Amici roof prism will erect the image and turn the optic axis through 45° or 90°. It is not recommended for amateur astronomy however, as the edge of the roof adds a faint streak of light to each star image.

2.5.3 The occasional astronomer

For visual observation of faint objects such as nebulae and for comet-hunting, the 7 mm exit pupil is the criterion for selection and not the magnification. It is a matter of personal judgement whether a large magnification is desired, but without a large and heavy objective there is little in the sky except the Moon and an occasional planet with sufficient brilliance to reward the observer. If faint objects such as nebulae are to be observed, the 7 mm exit pupil is mandatory. Probably the most satisfying instrument for an amateur astronomer and occasional observer is a pair of 10 × 70 or 12 × 80 binoculars. These will be heavy enough that a tripod with a pan-and-tilt head will be needed for a steady image.

Of course nothing is changeless. There are high refractive index plastics and doubtless someone in due course will produce a lightweight instrument with a $\times 15$ or $\times 20$ magnification with a 7 mm exit pupil. It will be portable but nevertheless will need a stand to give a steady image. It is worth mentioning in passing that a faint object like the spiral nebula M31 in Andromeda is almost a naked-eye object. A visual telescope or binocular with a 7 mm eye-pupil will give an enlarged view of it: but the view is still only of the central condensation. Only a long-exposure photograph will show the whole spiral, which has a major axis subtending about 3° on the sky. In other words, the object is about six times the size of the Sun or Moon, but alas too faint for the human eye to see.

2.5.4 Photography through the eyepiece

The above example of an exit pupil serves to illustrate a general principle: when two centred optical systems are coupled together, **the exit pupil of one must be the entry pupil of the other**. The image seen in a telescope can be photographed, but the eyepiece must be adjusted to put the observed image at $-\infty$ (which makes the telescope *afocal*) and the entry pupil of the lens must be at the exit pupil of the telescope. If the camera is of the single lens reflex (SLR) type or with a digital LCD viewfinder, the final adjustment can be made by moving the camera and observing the image, which will be unvignetted when the camera is at the correct position.

2.6 Applications of the refracting telescope

2.6.1 The theodolite

Theodolites date from the seventeenth century and are the classical instruments for surveying. The telescope is typically of high magnification with a small exit pupil. The image is usually (but not always) inverted and there is an eyepiece graticule [2] for directing the optic axis towards the feature to be measured. The important thing for the surveyor is the pair of divided circles which are the essence of the instrument, and the levels and screws which render the two rotation axes precisely horizontal and vertical.

2.6.2 The cathetometer

This a laboratory device for the accurate position-measurement of a close but non-contactable object. It comprises a simple Keplerian telescope designed to

[2] Considered in more detail in Chapter 3.

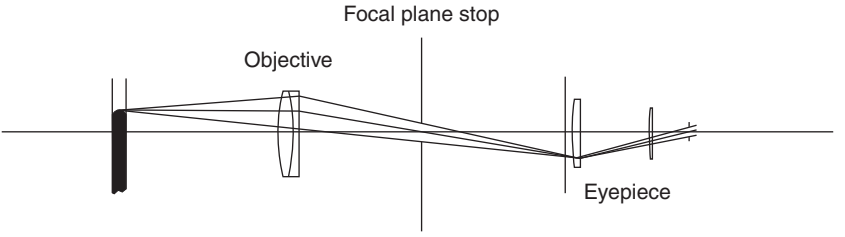

Figure 2.5 The cathetometer. This is a laboratory instrument, used chiefly for measuring the height of a mercury column in a barometer or manometer. The stop at the rear focal plane of the objective ensures that the instrument is *telecentric on the object side*, which ensures that the incoming chief-ray is always tangent to the top of the meniscus.

focus on a nearby object. It has a field stop at the rear focus of the objective (not at the image position, note), which renders the instrument telecentric on the object side. The combination ensures that all chief-rays from the object are parallel to the optic axis so that a true profile of a local three-dimensional object is observed. It is chiefly used in the laboratory to measure accurately the height of the meniscus of a mercury column. For this purpose it is usually mounted on a vertical graduated column with a ratchet for adjusting the height.

2.6.3 The profile projector

This is sometimes called a shadowgraph. It is in essence a cathetometer with a large diameter objective and again is telecentric on the object side. This allows the magnified projection on to a screen of the true silhouette [3] of an object under inspection. This in turn allows accurate measurements to be made of the dimensions of the object. It is essential that the objective lens or mirror diameter be larger than the largest object which is to be measured.

2.7 Reflecting telescopes

Focusing by reflection from a concave mirror originated with Isaac Newton (1642–1727) who was persuaded, wrongly, that all glasses dispersed light in the same way and that chromatic aberration was ineluctable. He also

[3] That is, the silhouette as seen by an observer at infinity. This is equivalent to the formal, drawing-office projection of the object.

discovered the algebraic geometry of the paraboloid and showed that all reflected rays coming from $-\infty$ and parallel to the axis of symmetry would necessarily pass through the same point, the 'focus' of the paraboloid.

2.7.1 The Newtonian telescope

Thus came the Newtonian reflecting telescope (of which Newton's second prototype is the library of the Royal Society in London). It has been improved over the subsequent centuries by Laurent Cassegrain (1629–1693), by James Gregory (1638–1675) and more recently and importantly by the collaboration of George Willis Ritchey (1864–1945) and Henri Chrétien (1879–1956). Practically all large astronomical reflecting telescopes are of the Ritchey–Chrétien type.

2.7.2 Properties

The focusing property of the paraboloid of revolution is matched by a similar property in the figures of revolution of some other conic sections. In the ellipsoid of revolution, rays emitted from one focus are reflected precisely to the other. In the hyperboloid of two sheets, rays directed towards one focus are reflected to the other. The Gregorian telescope makes use of the ellipsoid and the Cassegrain of the hyperboloid. Over the years the Cassegrain configuration has proved most fruitful because of the variations that are possible. The Gregorian however had the advantage that it produced an erect image, which made it suitable for naval or military use with simple eypieces.

2.8 The Newtonian telescope

Its chief merit is simplicity and the availability of a low focal ratio; Newtonians work typically at F/5 or thereabouts. However, if it is a large astronomical telescope the observing position is decidedly uncomfortable and possibly perilous when the observer is on a high platform in a completely darkened observatory.

2.9 The Cassegrain telescope

In this configuration the primary mirror is a paraboloid and the secondary mirror is a hyperboloid.

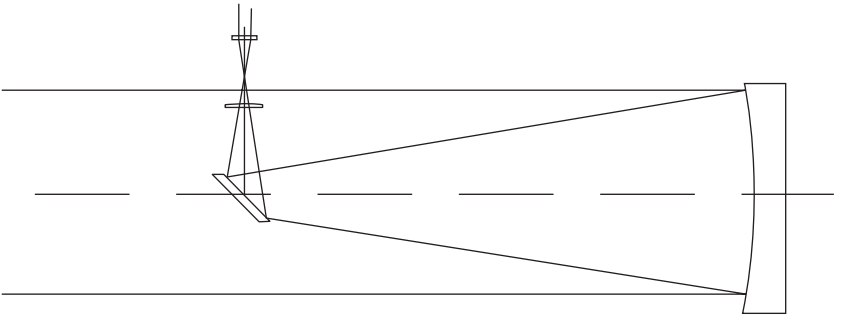

Figure 2.6 The Newtonian reflecting telescope. The secondary mirror needed to extract the image has a diameter typically one-third the diameter of the objective, and therefore obstructs about one-tenth of the light entering the system. In very large astronomical telescopes the secondary mirror-supporting structure is big enough to hold an observing station at what is called the 'prime focus'.

2.9.1 Variations on the Cassegrain configuration

The Ritchey–Chrétien and its relatives

The Ritchey–Chrétien telescope has a hyperboloidal primary, and the secondary can be chosen and figured to give freedom from spherical aberration and primary coma over a larger field than the classical Cassegrain.

There are other useful variations on the Cassegrain theme. The Dall–Kirkham configuration, for example, uses a spherical secondary and an ellipsoidal primary, and has the merit that it is simpler to figure and polish than a classical Cassegrain. It gives good correction over a limited field. It is usually used in small telescopes, but the 152 cm telescope at the Observatorio del Teide in Tenerife has this configuration. The configuration known as the three-mirror-anastigmat allows further corrections of the aberrations remaining in the Ritchey–Chrétien design, including a flat field, without needing refracting surfaces. The James Webb Space Telescope has this configuration.

The Maksutov Cassegrain

This design, by D. D. Maksutov, was first described in 1944[4] and was originally intended as a cheap, easy-to-make version of the Cassegrain telescope for use in schools. All the surfaces are spherical and the spherical aberration of the primary is corrected by using a meniscus lens (also with spherical surfaces) in front of it, with the secondary mirror a silvered patch on the rear surface of the corrector. It has the desired qualities but has a very small field of view and a high focal ratio (f/15 to f/22). There have been many variations

[4] D. D. Maksutov, New catadioptric meniscus systems. *J. Opt. Soc. Am.* **34**, 270–283 (1944).

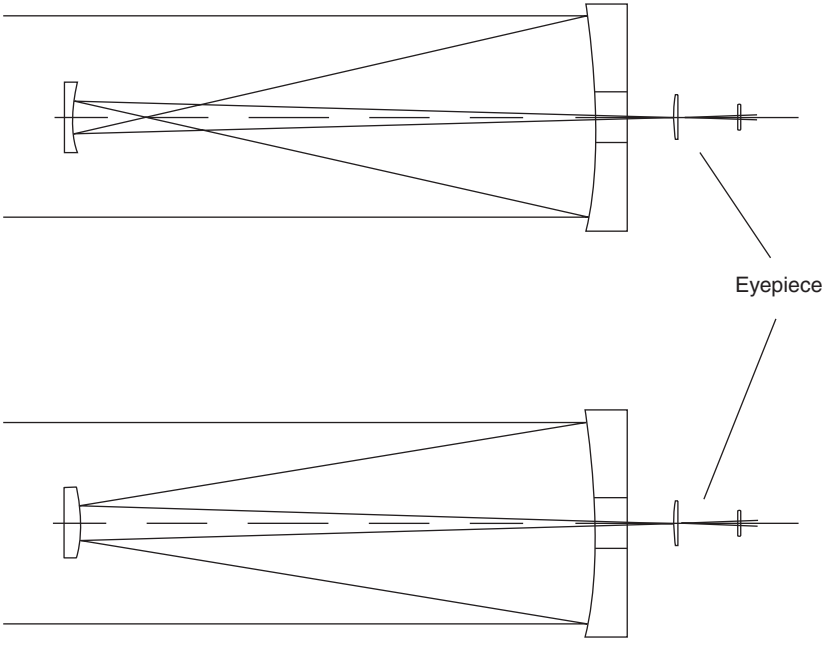

Figure 2.7 A Gregorian (top) and a Cassegrain (bottom) telescope. The two mirrors offer more opportunity for correcting aberrations (particularly coma) and various versions have hyperboloidal, spherical and parabolic primary mirrors, with secondary mirrors to compensate their aberrations. Chief among these is the Ritchey–Chrétien configuration of the Cassegrain, in which both primary and secondary mirrors are hyperboloids, giving a field of view of several degrees practically free from aberration. There is a considerable mechanical advantage as well, since heavy items of equipment such as spectrographs are attached behind the primary mirror, thus allowing a lightweight rigid support for the secondary mirror. Also worthy of consideration is the convenience and safety of the observer who, at the Newtonian focus of a large telescope, is sometimes in a precarious position near the top of the telescope tube.

on the theme with improvements both in quality of correction and in focal ratio. Basically it is a lightweight, high magnification telescope for advanced amateur astronomy and for remote inspection of engineering structures and buildings by maintainance crews. In connection with a cooled CCD camera it is used by military and law-enforcement agencies for clandestine observation in low light level conditions.

2.9.2 The Schmidt camera

This device was first described in 1929 by Bernhard Schmidt of the Hamburg Observatory and caused a major sensation in the astronomical world. It is

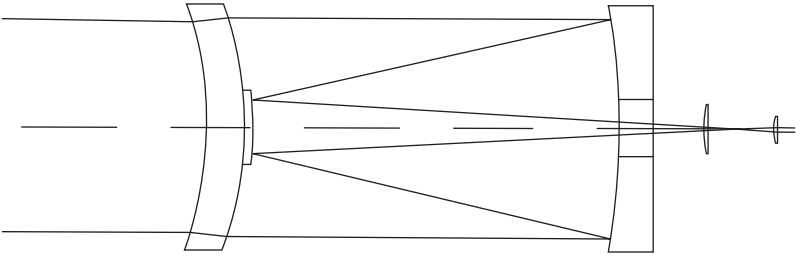

Figure 2.8 The Maksutov Cassegrain telescope. These are generally small astronomical telescopes of less than 50 cm aperture. The positive spherical aberration of the primary mirror is balanced by the negative spherical aberration of the meniscus corrector plate. All the optical surfaces are spherical, so that they are comparatively cheap to make, and have excellent aberration correction over a field up to 0.8°. They are much used by serious amateur astronomers as well as by the aeronautical industry for remote observing of aircraft and sounding rockets. There are also very long focus catadioptric 35 mm camera lenses of this configuration.

primarily of astronomical use and has the ability to photograph a much wider field than is available to Cassegrain telescopes, without significant aberration of the images. It comprises a spherical mirror with a 'corrector plate' at the centre of curvature (which is the entry pupil of the system). Because a spherical mirror has no unique optic axis (any line through the centre of curvature will do) there can be no 'off-axis' aberrations such as coma and astigmatism, and the corrector plate distorts the incoming plane wavefronts in the proper way to correct spherical aberration on reflection. The remaining aberration is field curvature and the photographic plates are bent over a mandrel, a machined spherical surface, to allow for this. The device allows photography of a wide field at a low F-number and consequently allows such photography of very faint objects.

It has a lesson for optical designers: if it is necessary to correct aberrations in a centred optical system, the corrector plate should be at a pupil of the system.

2.10 Refractors versus reflectors

For a given aperture:

a reflector is cheaper, shorter in tube-length and lighter in weight,
a refractor has a wider aberration-free field of view,

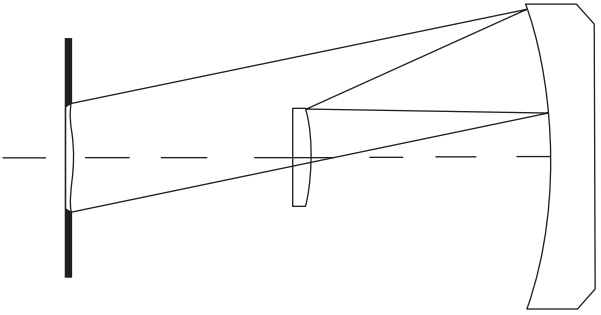

Figure 2.9 The Schmidt camera. This was a major sensation in the astronomical world when it first appeared, having an enormously increased numerical aperture and at the same time an enormously increased field of view. The chief difficulty of construction was the corrector plate at the left of the diagram, which until recent years was worked and polished by hand. Later models have chromatically compensated corrector plates of more elaborate surface form.

a refractor has about 20% higher light transmission,

a reflector has perfect chromatic aberration correction: a refractor often has a trace of 'lateral colour' especially if its lens is an 'astrographic' lens, an achromat corrected for blue-green/near ultra-violet light, made before the days of panchromatic photographic emulsions.

2.11 Focal reducers

This phrase 'focal reducer' is used to describe two different optical devices. In the simple case it is a positive aplanatic lens inserted before the Cassegrain focus of a telescope to increase its numerical aperture or, what amounts to the same thing, reduce the focal ratio or F-number, so as to produce a smaller, brighter image for photography. It gives only a limited field of view at best resolution. Alternatively, it is a more elaborate device, intended to be used on large astronomical telescopes and spectrographs, which is in effect a giant eyepiece, providing a collimated image and a real exit pupil with a diameter to fit the photographic lens of a CCD camera, typically a lens with an f/2 focal ratio.

The device was invented independently by G. Courtès of the Observatoire de Marseille and by A. B. Meinel of the University of Arizona. In essence it is a gigantic Kellner eyepiece followed by a 35 mm camera lens, and it allows astronomical photography at f/2.

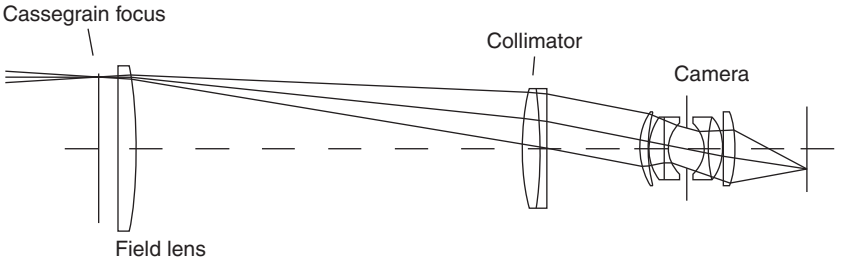

Figure 2.10 The focal reducer. This is, in effect, a giant eyepiece and converts a large astronomical telescope into a 35 mm camera lens. It chiefly allows 'deep' photography of very faint objects at a focal ratio of f/2 or better although, unlike the Schmidt camera, with a restricted field of view.

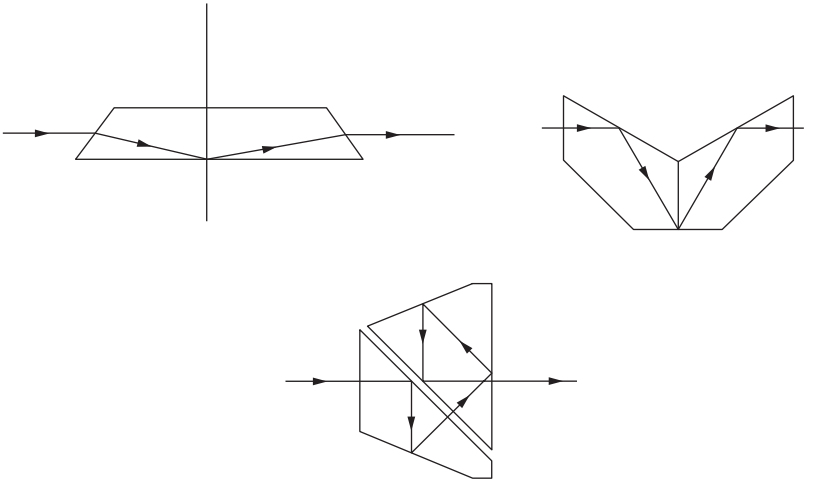

Figure 2.11 Image rotators. These are afocal systems for the end-on observation of rotating machinery, and on alt-azimuth mounted telescopes where prolonged observation involves rotation of an astronomical image. As such they are 'de-rotators', with their rotation angles guided, like the rest of the telescope movements, by a dedicated computer. They are generally similar to the image erectors of Figure 2.4 but with the odd number of reflections necessary to allow rotation. These prisms will *rotate* an image but will only *invert* it at the expense of having a mirror image of the object.

The field lens serves to direct light from all parts of the telescope field on to the pupil of the camera lens, i.e. it images the telescope objective on to the camera pupil. The eye lens, which may be an achromat, re-collimates the light from the Cassegrain focus so that the camera lens is adjusted for infinity.

2.12 Image rotators

These are devices which can be inserted into an image-forming centred system which will rotate the final image about the optic axis. Usually it is the rotator itself which is made to turn mechanically, moved either manually by the observer, or continuously by a motor, or automatically by a servo-mechanism. Rotators are used with large astronomical telescopes which are on alt-azimuth rather than equatorial mountings because of the need to 'de-rotate' the image as the telescope turns during the night. Rotators are used to inspect rotating machinery such as the fan-blades on gas turbines, to check the vibration patterns set up at certain rotation speeds. They are used to demonstrate the Coriolis force on rotating turntables when objects are made to move on them.

One of the oldest image rotators is the Dove prism. It has the sole advantage of simplicity and is adequate for simple demonstrations, but for serious use has several disdvantages. Firstly light is transmitted through a transparent medium with oblique incidence, and it has a single internal reflection with consequent inversion or mirror-imaging of the object being inspected. The optical aberrations of a transparent parallel-faced block of glass with oblique incidence at about 45° are severe, so the device can only be used in collimated light (the object effectively at $-\infty$) with small field angles.

Other rotators [5] avoid these disadvantages but sometimes add their own. For example the *Abbé rotator* of Figure 2.11 has normal incidence but is markedly asymmetric and with a long optical 'tunnel', which restricts the field of view that is available. For the inspection of rapidly rotating machinery the *Pechan rotator* is most appropriate. It has mechanical symmetry about the optic axis. It has normal incidence which minimizes (but does not totally eliminate) optical aberrations. Its chief demerit is that it mirror-inverts the image. [6]

[5] For a comparative survey of image rotators see D. W. Swift, *Opt. Laser Technol.* **4** (4), 175–185 (1972).

[6] Beware of confusion with the Schmidt–Pechan inverter, superficially similar to the Pechan rotator but with a roof prism to give an extra reflection which eliminates mirror-imaging. The Schmidt–Pechan prism of Figure 2.4 inverts **but will not rotate** the image.

3

Eyepieces, eyes and colour

3.1 Eyepieces

When you look through a telescope or binocular you focus the distant scene by first pulling out then slowly retracting the tube, or by turning the appropriate screw. For visual observation you should always extend the tube beyond the point of good focus and then draw it in until the scene is seen to be sharp and well focused. This is to avoid the eye strain which comes from a prolonged accommodation of the eye to a close object. [1] (The same is true when using a hand-lens. Bring the object in from a distance; do not start with it too close because, although it will be in focus well enough, the image will be close to the eye instead of at $-\infty$ where it belongs.)

The telescope field of view is bounded by a circular stop which is usually sharply in focus. There may be an eyepiece adjustment to make it so. It is a stop placed at the prime focus, and it is there to prevent scattered light from the outer, unusable part of the field from entering the eye, thereby reducing the contrast of the scene. A second such stop, with the same purpose, is at the intermediate pupil of an erecting telescope. The two stops are optically conjugate.

The diameter of the stop may be determined by the limit of good focus of the field, because the off-axis aberrations of the objective gradually increase with field angle. That, however, is very unlikely: much more important is the limit of sharp focus imposed by the eyepiece itself since there is likely to be both astigmatism and chromatic aberration in the outer field and these become intolerable when the field angle is big enough.

Eyepieces were invented at the same time as the Keplerian telescope and two in particular outshone all others in performance in the early days of the

[1] Ancient observers like the present author no longer suffer from this problem, having long ago lost the accommodation which the young eye possesses.

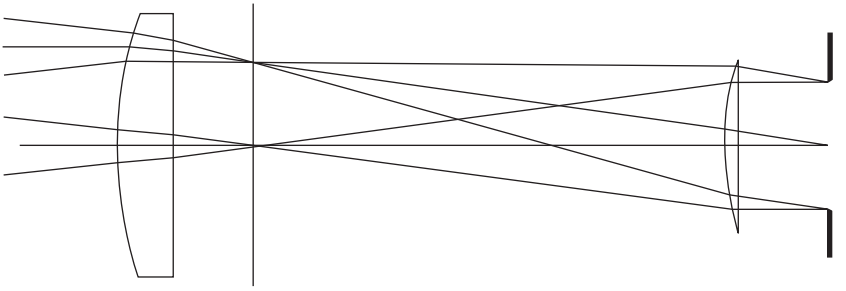

Figure 3.1 The Huygens eyepiece. This was probably the earliest device designed to improve the field of view of early Keplerian telescopes and at the same time, by requiring that the separation be half the sum of the focal lengths, it reduced chromatic aberration. It is now obsolete because of its limited eye relief and its inability to correct the aberrations of modern short focal length objectives.

Enlightenment (the seventeenth and eighteenth centuries). The first satisfactory one was the invention of Huygens in ∼1670, used until it was surpassed by one invented by Jesse Ramsden (1735–1800) in 1780.

The Huygenian eyepiece comprised two plano-convex lenses of the same glass, of nearly equal focal length, separated by a distance equal to half the sum of their focal lengths. This separation ensured that the focal length was independent of the variation of refractive index with wavelegnth so that there was no lateral colour. On the other hand the eye relief of the arrangement is necessarily very small and the field lens comes before the primary image so that there can be no external graticule.

As mentioned in Chapter 2, the purpose of an eyepiece is twofold.

(1) To re-collimate the light which has been brought to a focus by the objective, so that there is a virtual image at $-\infty$ of the real image formed by the objective. In this sense the eyepiece is equivalent to the watchmaker's lens which is used to examine a close fine object.
(2) To form a real image of the objective on the observer's eye-pupil. Then this image, the exit pupil of the first part of the system (the telescope), coincides with the entry pupil of the next part (the observer's eye). As in any well-designed optical system, this ensures that all the light which enters the objective arrives on the retina of the observer or at the photographic plate or CCD or any other device which is the final detector of the light.

Since those days there have been improvements in the design of eyepieces, with several aims, not all achieved at the same time. A wide field of view is useful in a telescope or a binocular if a good observation is to be kept on an area

sparsely populated with objects of interest. High resolution at the centre of the field is desirable for the close examination of fine details – for example under the microscope. Freedom from chromatic aberration is always desirable. A flat field is desirable if the eye is unable to accommodate rapidly when rotated to look at the edge of the field (this in an eyepiece with a large exit pupil, designed especially for the purpose). A constant magnification (i.e. zero distortion) is desirable across the field to prevent odd 'bulging' effects as the telescope is moved. The question of internal re-imaging by internal reflection arises, but is much abated with anti-reflection coatings on the glass-air surfaces.

3.2 Eyepiece properties

For the most part telescopes and microscopes come with an eyepiece already installed, but in constructing an optical system it may be possible, indeed necessary, to make a choice. Eyepieces are made with standard barrel-diameters, so that they can be changed relatively easily. [2]

The criteria which affect the choice are the following.

(1) The visual resolution – which is almost always adequate.
(2) The *apparent* field of view, the angle subtended by the edge of the field stop at the pupil of the observer's eye.
(3) The eye-relief, the distance from the last glass surface to the exit pupil – the place for the observer's or camera's eye-pupil.
(4) The distortion, the regularity of a square mesh grid on a microscope table for instance. This is of comparatively low importance unless measurements are being made. The eye–brain combination is very tolerant of distortion in an image.
(5) The field-curvature, the ability to focus the eye on any part of the field of view without requiring accommodation.

The list of designs which achieve most (but rarely all) of these qualities is long. What follows is a short account of some of the favourites among users.

3.2.1 The Ramsden

This is the classic design with two plano-convex lenses, the convex surfaces facing inward toward each other. It is still in use, sometimes with polystyrene lenses, in simple telescopes where no exceptional performance is expected.

[2] But be careful: there are several standards, not all compatible.

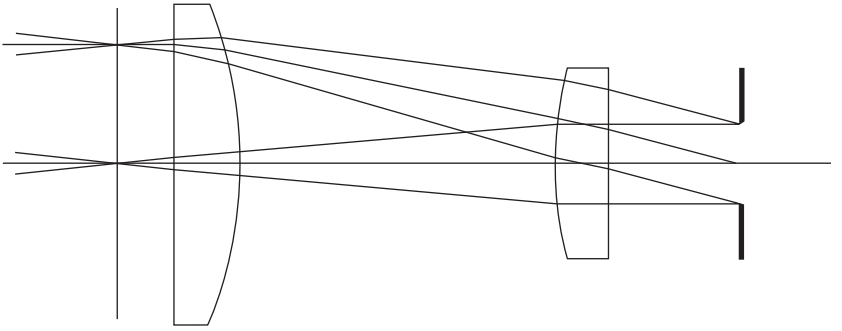

Figure 3.2 The Ramsden eyepiece. This was a distinct improvement on the Huygens but almost a century later in its invention. It had improved eye relief and as the forward (f_1) focus was before the field lens, spider-silk cross-hairs could be added to give fiducial marks for astrometry and time-keeping.

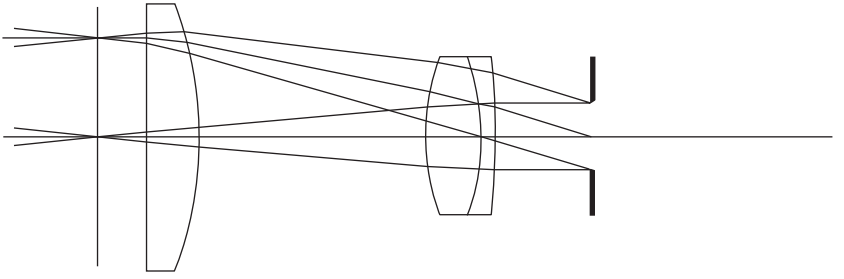

Figure 3.3 The Kellner eyepiece. The chromatic aberration correction of the Ramsden eyepiece was incomplete and the achromatic doublet was still in the process of invention in Ramsden's time, so that the obvious improvement needed was introduced by Carl Kellner (1826–1855), a German optician and lens-maker. Kellner eyepieces are still in use.

3.2.2 The Kellner

The Kellner [3] eyepiece is a Ramsden eyepiece with the single plano-convex eye lens replaced by an achromatic doublet. It is simple, inexpensive and adequate for small [4] telescopes where no exceptional performance is required.

3.2.3 The Abbé or orthoscopic

This has ample eye-relief and a distortion-free image (which is why it is called 'orthoscopic') but has a comparatively small field of view.

[3] Carl Kellner (1826–1855), founder of the Leitz company.
[4] 'Small' here means with objectives less than 75 mm aperture.

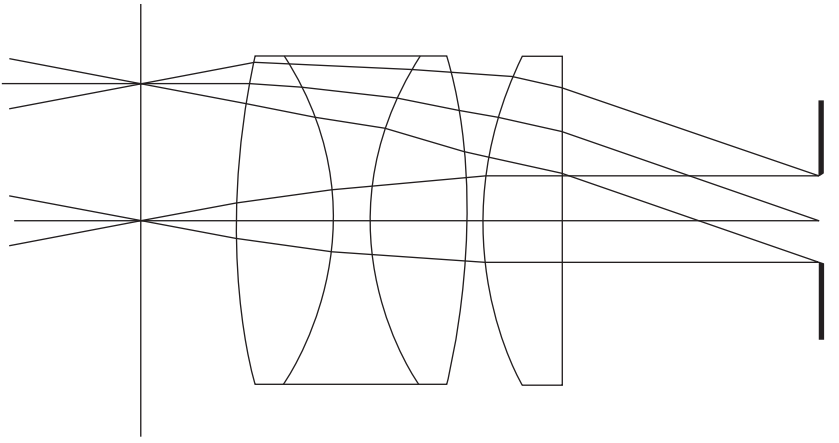

Figure 3.4 The Abbé orthoscopic eyepiece. There was a serious reconsideration of the principles of eyepiece design by Ernst Abbé (1840-1905) of the University of Jena, the founder of modern microscopy and co-owner of the Zeiss optical firm. His eyepiece is free from distortion and has excellent eye relief for spectacle wearers. As it was intended for microscopy, its field of view is limited – perhaps too much for visual astronomy.

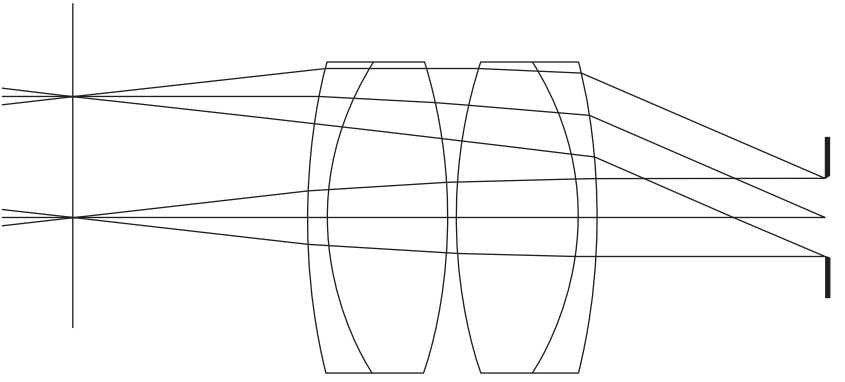

Figure 3.5 The Plössl eyepiece. This was forgotten for a century but has since become a favourite with amateur astronomers for its wide apparent field of view. Again it is limited by its short eye relief which makes it inconvenient for spectacle wearers.

3.2.4 The Plössl

This has a wide field of view which has made it popular with amateur astronomers, but has a short eye-relief, which makes it unsuitable for people wearing spectacles.

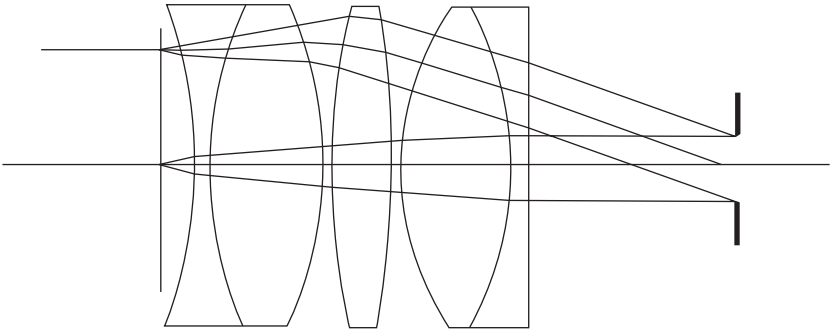

Figure 3.6 The Erfle eyepiece. This comes close to the ideal for many people and is probably the first choice for amateur astronomers. The field of view, freedom from distortion and eye relief are ample. It was chosen for naval rangefinders because of these properties and because of the comfort of viewing afforded by the large exit-pupil in comparatively low magnification telescopes.

3.2.5 The Erfle

This has both an excellent field of view and adequate eye-relief. There is a possible problem: the forward focal plane, and hence the graticule position, is very close to the front surface of the field lens so that dust on this surface, albeit slightly out of focus, may be an irritating problem. This type of eyepiece comes close to the ideal and was much used, for example, in naval rangefinders. It is a favourite among serious amateur [5] astronomers, particularly among comet-hunters.

3.3 Astronomical photography through an eyepiece

The camera iris must coincide with the exit pupil of the telescope (check that there is adequate eye-relief and that the pupil diameters are the same). An opaque tube should surround the light path and fill the gap caused by the eye-relief. The only purpose of using a camera this way is to improve the large focal ratio available by using a Cassegrain telescope itself as the camera lens, so as to reduce the exposure time on faint objects. **This only works if the diameter of the telescope exit pupil is the same as the iris diameter of the camera lens.** Consequently a low power eyepice is indicated since the exit pupil diameter is the objective diameter divided by the magnification.

[5] And some professionals, including the present author.

3.4 Visual observation

This is regrettably rare these days among professional astronomers: telescope time is too valuable.[6] But be careful not to use too *low* a magnification with a large reflecting telescope. The exit pupil is the image of the objective mirror and the secondary mirror, whether Newtonian or Cassegrain, is superimposed upon it. With a low magnification the exit pupil is large and may be large enough that the central obstruction has the same diameter as the eye-pupil, in which case no light will enter the eye.[7] In visual observation, the ideal always is to match the pupil diameters so that the image of the eye-pupil fills the aperture of the telescope objective and *vice versa*. Atmospheric turbulence ('seeing') limits resolution to 1 second of arc, so that there is no point in trying for more than $\times 60$ magnification.

3.4.1 Eyepiece graticules

Graticules, or cross-hairs, or fiducial lines, were originally included in the eyepieces of astronomical instruments, and arranged to be in the focal plane of the objective (and the eyepiece). Traditionally spider-web silk was the preferred filament, and the eighteenth century instrument maker Jesse Ramsden bred a species of garden spider (*Araneus diadematus*) specially for the purpose. The filaments were easily broken and, for replacement, short lengths, stretched across a card with a U-shaped cut-out were packed in boxes for transport. Alternatives were silica threads, traditionally formed by heating a silica rod to melting point and firing one end from a cross-bow, catching the thread on a black velvet cloth. They were illuminated for astronomical work by a candle flame in a side-tube.

More recently, they have been made by photographing fine lines and patterns (with numerals) on a high resolution emulsion or by physically scratching lines on to glass discs. This latter method was particularly useful, as the scratches could be illuminated easily by the scattering of light fed in from the side.

For observing an edge (e.g. a spectrum line with the measurement to be made on one edge of the image of the entry slit) the eyepiece should have a pair of mutually perpendicular cross-hairs. They should be rotated to be at 45° to the edge.

[6] It is estimated that the cost *per night*, cloudy or fine, of a large research telescope is the same as the observatory director's monthly salary.

[7] Some amateur astronomical societies have telescopes with 450 mm aperture or more. With a $\times 25$ eyepiece the exit pupil diameter will be 18 mm and the central obstruction image diameter will be about one-third of this, or 6 mm diameter. This is big enough to exclude most of the telescope light from the eye-pupil.

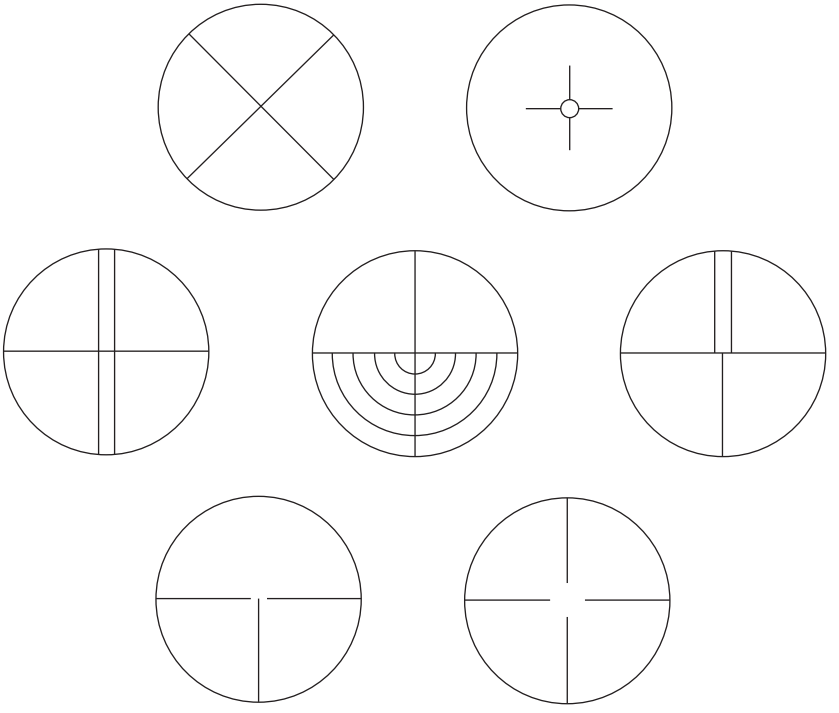

Figure 3.7 Eyepiece graticules. These are designed for a variety of purposes and can be adapted to the physiology of the eye. Not all observers have high resolution at the fovea, and parallel lines rather than a single vertical line are sometimes more convenient and more accurate for the measurement of horizontal angles with a theodolite. This is particularly true in low light level measurements.

When using a telescope to look at a distant vertical object (a mast, a tree, a lighthouse), a graticule with a pair of parallel vertical hairs should be used, since the eye–brain combination can judge an inequality in spacing to an accuracy well beyond the eye's visual acuity, and with a fine screw the eye can place a distant vertical object to 1 second of arc midway between the two parallel graticule lines. This is a much smaller angle than can be resolved at 5 km distance, where again atmospheric turbulence limits visual resolution to about 10 seconds of arc. In celestial navigation, the disc of the Sun can be most accurately centred between two parallel hairs separated by a distance slightly less than the disc diameter.

If the graticule includes a scale for the measurement of small angles (or distances in a microscope), be quite certain before beginning that there is no parallax between the object and the scale. Both should be imaged at infinity.

If the eyepiece has distortion the graticule may appear non-linear. This must be ignored and the graticule scale trusted.

For visual astronomical work the graticule must be illuminated. For stellar positions [8] a circle subtending about 4 seconds of arc on the sky is best, illuminated with green light. The brightness should be adjustable and interrupted, on-off equiphase, with a period of about half a second. This is because, with very faint stars, the light from the graticule tends to make the image disappear if the light is continuous.

3.5 The eye as an optical instrument

This is not the place for a thorough discussion of the physiology of the eye, but there are a few interesting remarks worth making.

Normally in daylight the eye works at about f/8. Visual acuity in the young eye is at best about 1 minute of arc and diminishes with age. For example, the planet Jupiter at opposition subtends at most 51 seconds of arc, so that it just, but only just, presents a disc and more often than not requires the eye of faith to make it so. In a $\times 7$ binocular, Jupiter is an obvious disc and his four brightest satellites can also usually be seen. Venus likewise presents a moon-like crescent at maximum brightness.

The eye however is organized differently by the brain and mere visual acuity is not its strongest attribute. In combination with the brain it can detect a lateral break or misalignment of a fine line to much less than 1 minute of arc, possibly because, surprisingly, the eye is vibrating rapidly (between 30 and 80 Hz) with an amplitude of about 20 seconds of arc – about the right amount to alternate an image point between two adjacent retinal receptors. It suggests that focusing is servo-controlled by the brain using an a/c signal and a phase sensitive detector. Ditchburn's seminal papers [9] on this topic are from as far back as 1952. He also found that an image, artificially stabilized on the retina, disappears after less than 1 second, indicating that the eye works by a/c signal transmission.

3.5.1 Photopic and scotopic vision

The eye is not designed to be a camera. It has evolved and is intended to track motion, to recognize natural features such as edges and corners and to

[8] Rarely measured visually, usually measured on a glass photographic plate with a travelling microscope.

[9] For example, R. W. Ditchburn and B. L. Ginsborg, *J. Physiol.* **119,** 1–17 (1953). See also L. A. Riggs and F. Ratcliff, *J. Opt. Soc. Am.* **42**, 872–873 (1952).

give sharp visual discrimination over a small area of the retina centred on the fovea. Regarded strictly as an image-forming optical instrument it is of remarkably low quality. As an information-processing device it is superb and well matched to the brain's information bandwidth of about 14 baud. The so-called rod and cone cells, the receptor cells of the retina, fulfil different functions. The cones cluster strongly around the fovea and give high spatial definition on a circle about 1.5° diameter. They are pigmented to give colour sensitivity. Beyond this circle the spatial density falls and there are essentially no cone cells beyond 25° from the fovea.[10] Rod cells are absent from the foveal region. They are more sensitive than the cones but have no colour discrimination, so that there is poor colour vision in the visual field beyond this 25° circle. Rod cells allow *scotopic* vision, that is so say colour-insensitive vision at very low light levels, with a sensitivity not far above the quantum-limited level. This gives a dynamic range to the eye of about 10^8 so that the faintest objects visible are one hundred-millionth of the brightness of an outdoor scene on a sunny day. It can be deduced that it takes typically eight photoelectrons to stimulate a rod cell. Because there are so few rod cells in the foveal region there is no fine spatial resolution so that sharp edges, for example, are not easily seen on a moonless night. 'Night glasses' as they are called (telescopes with 7 mm exit pupils) magnify images sufficiently that shapes and edges are more easily distinguished. There is no increase in surface brightness (which would be physically impossible) but there is a gain in visual acuity over the unaided eye, limited by the photon shot-noise in the vision system.

3.5.2 Purkinje effect

Associated with scotopic or night vision is the *Purkinje effect*, first described in the early nineteenth century by the Bohemian scientist J. E. Purkyně, in which the maximum of visual colour sensitivity appears to shift towards the blue in faint light.

3.6 The perception of colour

A rational account of the way we distinguish between different regions of the visible spectrum was given by Thomas Young (1773–1829), who deduced that

[10] Which is perhaps why your TV screen should subtend about 50° from your eyes.

there were at least three different kinds of sensor on the retina, sensitive to different wavelength regions in the visible spectrum. Clerk Maxwell, after experiments with spinning tops with coloured segments, suggested that the colours we see could be synthesized from three 'primary colours', choosing red, green and blue as these primaries. Later it was found that by superimposing light with appropriate intensity from three projectors of the three primaries, white light will appear. Later still it was found that colour vision is more complicated than this; that some information processing goes on in the retina, as indicated by the fact that the number of nerve fibres connecting the eye to the visual cortex is far fewer than the number of retinal receptors, and that some rare forms of colour-blindness, such as axanthopsia, indicate a failure of the processing mechanism.

Maxwell's synthetic white is not the same white as the light from a radiator such as the Sun, and the spectroscope instantly shows the difference by demonstrating 'pure' colours, in other words the effect that monochromatic radiation produces on the retina and thence on the visual cortex of the brain.

Colour in practice is measured by the *colour triangle*, a practical application of homogeneous coordinate geometry. Points inside a triangle are joined to the vertices by three lines and the areas of the three subtriangles are three coordinates, each representing the 'strength' of its opposite vertex. If the three vertices denote red, green and blue lights, a point inside the triangle measures the proportion of each in any observed coloured light, 'white' being defined as the point where the three triangles have equal areas. For example, a point in the triangle near the centre but displaced towards the red vertex will indicate a shade of pink, an 'unsaturated' red.

Pure spectral colours cannot be represented by points inside the triangle, and an arc can be drawn around it to represent them. For example, the purest yellow which we can produce with three coloured lights is an unsaturated yellow, somewhere between the pure colour and white. There is no unique shape to the arc, as it depends on the particular colours or dyes which are chosen to be at the corners of the triangle.

All this is, in any case, an approximation, because in practice the three primary colours are not monochromatic, but – except for laser light – consist of fairly broad bands of the spectrum which have passed through three coloured filters. Both colour photography and colour printing rely on the fact that colours in Nature tend to be highly unsaturated, so that colour photographs can give an acceptable (though often not very accurate) representation of the original scene – and incidentally at a higher saturation than the original because a truly accurate reproduction of Nature would appear somewhat disagreeably 'washed-out'.

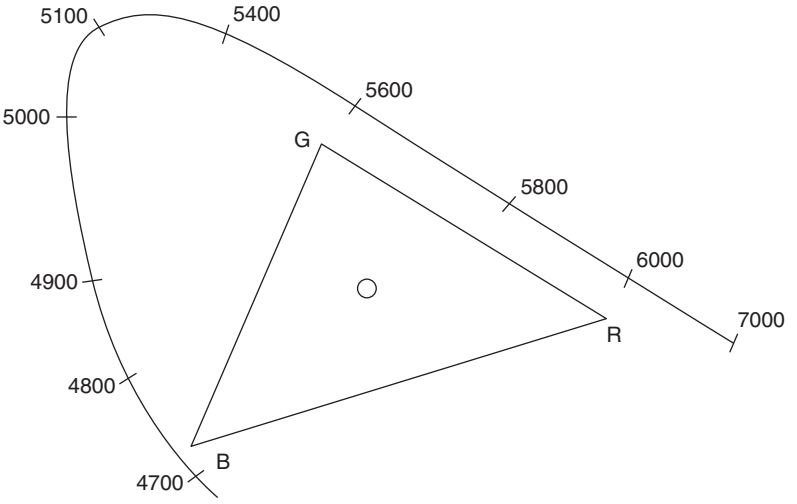

Figure 3.8 The colour triangle. The actual shape depends on the particular standard colour filters or light sources chosen for the red, green and blue vertices. The surrounding curve, on which wavelengths are marked in angstrom units, represents saturated or 'pure' (i.e. spectroscopic) colours. Only colours within the triangle can be reproduced in photographic prints or diapositives, although most colours in nature are very unsaturated and can be found within a small area around the 'white' spot at the centre of the triangle.

The arc representing pure spectral colours does not close. The implication is that the eye can see colours which are not in the spectrum. In fact all the red-blue mixtures such as purple, puce and magenta are non-spectral colours.

Likewise, there is no absolute 'white' and one is at liberty to define it. It may be taken as the colour of sunlight, or light in which all wavelengths are present in equal amounts (i.e. with equal power per unit wavelegnth range or wavenumber range). A white surface, on the other hand, can be defined: it is one which reflects and scatters all luminous energy which it receives. A coloured surface is one which abstracts selectively from different spectral regions. This in turn gives an interesting conundrum. There are objects which we call 'brown': yet there is no 'brown' light.

A black surface is, by the same criterion, one which absorbs all the luminous energy which it receives. Neither black nor white surfaces exist in Nature. There are various shades of very dark colour ('low albedo', as the astronomers call it) and there are what artists and decorators call 'warm' and 'cool' whites, very unsaturated reds and blues. 'Brown' in fact is yellow or orange with a very low albedo. An experiment in a dark room with a white spotlight and cards of various colours will demonstrate this.

3.6.1 Colour television

The sensation of colour is produced on a television (or any VDU) screen by red, blue and green self-luminous pixels so closely spaced that at a normal viewing distance the eye cannot resolve them.

3.6.2 Colour printing

Both diapositive colour slides and the inks used for colour printing rely on subtracting colours from white light, so that what remains is what we see. A *cyan* ink or dye absorbs red light and transmits green and blue light. *Yellow* ink absorbs blue light and transmits red and green. *Magenta* ink removes green light and transmits red and blue. Thus, for example, a cyan filter superimposed on a yellow filter will absorb firstly red light, then blue light, allowing only green light to pass or to be reflected. Cyan and yellow inks superimposed on white paper similarly will reflect only green light. Nevertheless, only the most basic colour printing will use just three inks. In practice as many as seven are required to give the subtlety of fine colour printing.

4

Cameras and camera lenses

4.1 Introduction: the photographic lens

A simple plano-convex lens has one surface plane and the other spherical. Its focal surface is also approximately spherical. The image of a point source at $-\infty$ on the optic axis is never a point because of spherical aberration, and the image is more and more distorted by other aberrations as the field angle increases.

Eighteenth century telescope makers such as John Dollond had discovered these aberrations long before the first photo-sensitive chemical materials were discovered, and they knew that both chromatic aberration and spherical aberration could be corrected (i.e. reduced to a satisfactorily small level) by combining two lenses with spherical surfaces, a positive one made of crown glass and a negative of flint glass, into an *achromatic doublet*, usually but not always a *cemented* doublet, in which the second and third surfaces were complementary and the two were joined by a thin film of Canada balsam. This was the first step on a long road which has led to the wide-aperture zoom lenses of current cameras.

A significant milepost on this road was the Gauss achromat, a positive meniscus followed by a negative meniscus: an air-spaced doublet which by its 'bendings', i.e. changing the shapes of the elements without changing their powers, reduced the aberrations. It was originally designed by C. F. Gauss at the request of Repsold the Hamburg lens maker, who needed an achromatic telescope objective of improved performance. The next brilliant step was by Alvan Clark of Bausch & Lomb, who put two of these back-to-back to get a short focal length camera lens. Later developments added extra elements to reduce aberrations even further. The result was the 'double Gauss' lens, a configuration which eventually replaced the 'Tessar' as the ultimate fixed-focus camera lens.

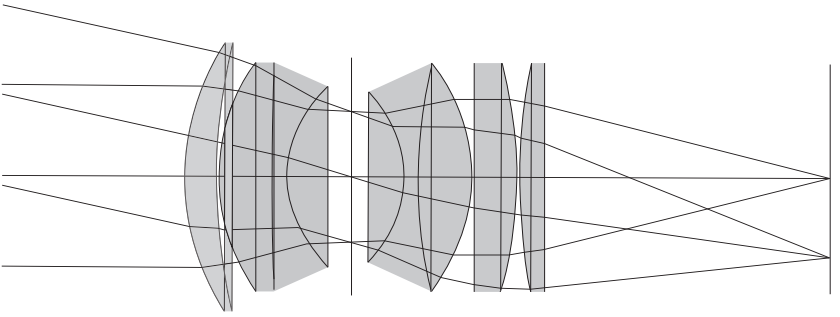

Figure 4.1 The double Gauss lens. This is one of the great achievements of the lens designer's art. It does something totally unnatural, in requiring rays of three different colours to arrive at the same point on the image, coming from the same direction at all wavelengths, and to do this from a wide variety of directions in the field of view on to a focal surface which is plane, and without distorting the image. It was invented early in the twentieth century, long before the days of the digital computer, when the lens designer's essential tool was a thick, heavy book of seven-figure logarithm and trigonometric tables.

4.2 The double Gauss lens

Figure 4.1 is the optical ray-trace diagram of a classic double Gauss design, with a focal length of 50 mm and a focal ratio F/2. These lenses first appeared about 100 years ago, and before the appearance of anti-reflection coatings were as far as it was possible to go in complication. The elements of the lens are of various glasses with different refractive indices and different dispersions (or v-values). It is designed to have the following properties.

(1) It must bring to a common focal point on a plane surface, light of three different colours in the visible region, coming from a point source anywhere in the distant object field. Furthermore, light of other visible wavelengths must pass through a circle centred on this focal point, of not more than about 2 or 3 μm diameter.

(2) It must do all this at a focal ratio of F/2 or thereabouts [1] over an angular field diameter of typically 48°. In other words, light rays of all visible wavelengths entering at a point at the first surface and taking various different wavelength-dependent routes through the interior, must emerge at the last surface, all at the same point and going in the same direction. To

[1] Lower focal ratios are achieved of course, but not at the strict limits given here.

arrange this is not easy[2] because it is asking a system of lenses to do something completely unnatural.

The path to the extraordinary degree of excellence of modern lenses has been long and thorny, marked by milestones such as the Petzval Portrait Lens (1840), the Rapid Rectilinear (1866), the Cooke Triplet (1893) and the Tessar (1902). It has been well documented by Rudolf Kingslake.[3] Since the mid twentieth century the important factor of variable focal length has been successfully added, making the achievement even more stupendous. The world of the lens designer is a small one and the modern high aperture zoom lens, taken for granted by the camera-owning public, is a source of almost miraculous ingenuity and invention within that world.

In the example shown in Figure 4.1, the lens designer has about 50 variables at his disposal (10 surface vertices, 10 curvatures, 15 refractive indices (typically) and 15 partial dispersions or v-values; (see Appendix B for details of these). I leave to the reader's imagination the thought of solving 50 simultaneous, non-linear differential equations. Of course lens design is not done that way and experience, intuition, invention, ingenuity and prodigious computer ray-tracing all play their part.

It is worth mentioning for would-be infra-red photographers, that the high degre of correction is rapidly lost beyond the visible region and photographic lenses generally do not work well outside their design parameters.

4.3 Telephoto lenses

The object of these lenses are to attain a large focal length without the accompanying long barrel which would be necessary in an ordinary lens. The basic principle is to follow one group of elements of comparatively short focal length with a second group with negative focal length. The principal planes of the combination are then substantially to the left (i.e. on the incoming or object side) of the first vertex and this ensures the required focal length of the assembly.

4.4 Wide-angle lenses

While lenses of focal lengths down to 21 mm have been designed for the usual 35 mm format, the problems of distortion and field curvature become

[2] A masterpiece of understatement.

[3] R. Kingslake, *A History of the Photographic Lens* (Academic Press, San Diego, CA, 1989).

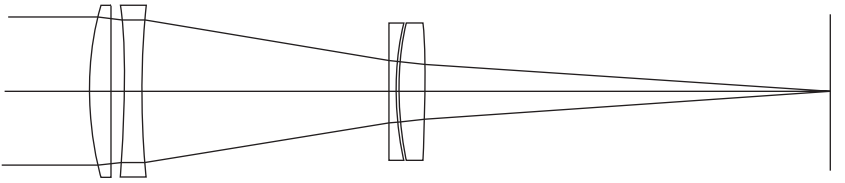

Figure 4.2 The telephoto lens. This is an adaptation of the Galilean telescope and gives a high magnification with a narrow field of view. Although the term is used indiscriminately for any long-focus camera lens, a telephoto lens, strictly speaking, is defined to be a positive lens-group followed by a negative group, intended to give a long focal length in a short barrel.

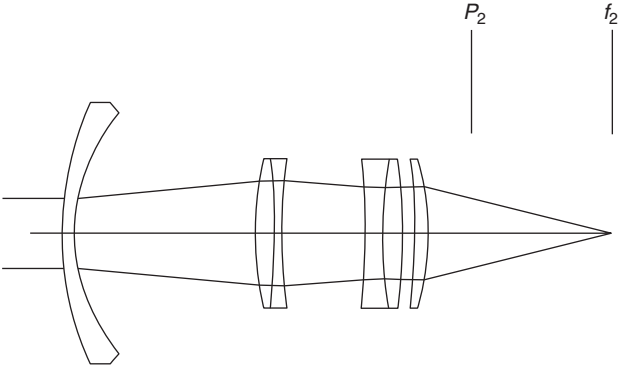

Figure 4.3 The inverted telephoto. This is a photographic device to allow a wide field of view without intruding into the camera body, and to avoid interfering with a reflex mirror or a focal plane shutter mechanism. It achieves this by putting the second principal plane, P_2, after the last glass surface. Such lenses, being highly unsymmetrical, tend to suffer barrel distortion, which may make them unsuitable for cinematography.

rapidly more severe and the trick then is to design a lens with an inverted telephoto layout. The first group, sometimes a single element, has a negative focal length and is followed (very closely in practice) with a second, positive group. Together they bring the principal planes closer to the focal plane than the lens itself. This has the advantage in SLR cameras that the reflex mirror is not obstructed by the rear element of the lens although the principal plane is just 21 mm from the focal plane. Such short focal lengths tend to have noticeable barrel distortion, tolerable in a single photograph, but disturbing in a moving picture when the scene is panned. Features which were small on the margin become enlarged as they approach the centre of the frame and then diminish

again as the panning continues. Symmetrical lenses (such as the double Gauss), free from distortion, are much to be preferred for cinematography. [4]

4.5 Fish-eye lenses

These are lenses of exceptionally short focal length and large field of view, amounting often to a field angle of 90°. They are so-called because of the view a fish might have of the sky above the surface of calm water, bearing in mind that rays almost parallel to the surface will be refracted at an angle of 48° to the vertical so that a fish would see the whole celestial hemisphere within a cone of 48° apex angle. R. W. Wood [5] took photographs with a pin-hole camera set vertically and filled with water to achieve photographs covering almost the whole hemisphere.

Aeronomers monitoring the night sky for meteorite trails have used standard cameras pointing vertically downwards towards a silvered glass sphere which reflects the whole area of sky (except obviously for a small area about the zenith covered by the camera itself). Several lens designers have produced extra-wide-angle lenses with a similar object, and outstanding among these are designs of Miyamoto Kenryo, [6] who departed radically from previous theory by using two huge, negative-meniscus elements to concentrate the light before starting the serious aberration reduction. He thus achieved focal lengths as low as 8 mm on the 35 mm format. The usual parameters used to describe a lens are abandoned. There are no fixed principal planes or pupils for instance, [7] and the exit pupil of the system depends on the field angle (which is why, in common with other short-focus lenses, the entry pupil appears to follow you uncannily like an accusing eye as you move sideways away from the camera). The short focal length is preserved over the whole field and solid angle is preserved, more or less, [8] over the whole frame.

These lenses project the celestial hemisphere on to a plane surface and thus they perform the same function as a map projection. A normal long-focus lens,

[4] This may be why the Taylor-Hobson 2-inch anastigmat, a double Gauss with a focal ratio of f/2, was the favourite lens of Hollywood film makers until the early 1950s.

[5] R. W. Wood, *Physical Optics* (Macmillan, New York, 1911), p. 68.

[6] K. Miyamoto, *J. Opt. Soc. Am.* **54**, 1060–1061 (1964).

[7] It is worth remarking that in the Miyamoto design the exiting chief-rays are nearly parallel to the optic axis, and the addition of a weak positive element at the rear will render them accurately parallel so that, with an interference filter following, it is possible to do monochromatic photometry simultaneously over the entire visible sky, isolating any one of the bright emission lines of the night airglow.

[8] A globe imaged as a disc measuring 5.0 mm at the centre of a frame, was imaged as an ellipse 5.0 mm radially and 5.4 mm tangentially in the corner.

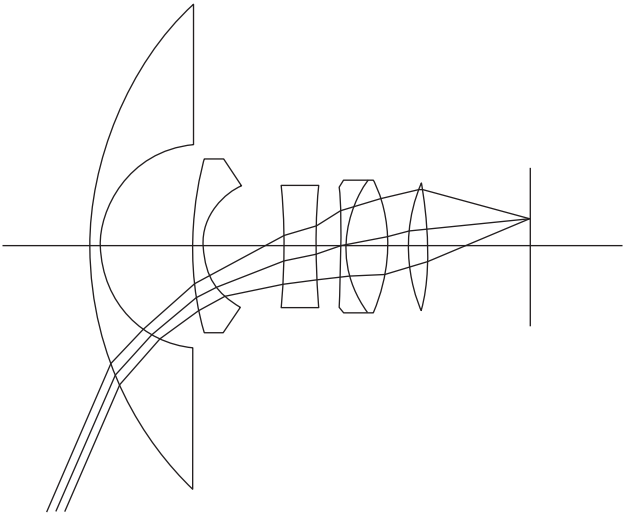

Figure 4.4 The Miyamoto fish-eye lens. This is the ultimate wide-angle lens, intended to photograph a complete hemisphere on to a plane surface. Apart from specialist artistic photography, its main application is in aeronomy, for photographing the night sky for meteor-trail studies or for airglow measurements. Despite its restricted usefulness, it remains one of the high peaks of the lens designer's art.

provided it is free from distortion (which most modern lenses are), produces a *gnomonic* projection, where the distance r of the image of a distant object from the centre of the picture is proportional to the tangent of the angle which the object makes with the optic axis of the camera:

$$r = f \tan\theta$$

Wide-angle lenses on the other hand may, depending on their design, produce projections closer to stereographic, in which case

$$r = f \tan\theta/2$$

and the ideal fish-eye lenses, known as 'R-theta' lenses will produce an image in which the distance from the centre of the photograph is proportional to the field angle on the object:

$$r = f\theta$$

In practice they come close to this ideal. What is important for photometry is that a lens should preserve *solid angle* so that the image of a small globe, for example, although circular when it is on the optic axis, is elliptical if it is in the

corner, *but with the same area* on the image. Nevertheless accurate photometry requires that any lens/CCD combination be calibrated before being used.

4.6 Zoom lenses

There is a long history of vari-focal lenses reaching back to Thomas Dallmeyer [9] in 1890. These are lenses with a changeable focal length but which must be re-focused before they can be used. 'Zooming in' on an object while filming was never possible, therefore, until the modern fixed-focal-plane zoom lens was devised by Harold Hopkins (1919–1994). He apparently realized the need while seeing BBC TV cameras with their heavy turrets of lenses at Lord's Cricket Ground in 1946. The number of zoom lens designs is legion and the designs are of wonderful complexity. Their working depends on using several groups of lenses with their relative positions controlled by cams as the lens barrel is rotated. Zoom ratios greater than 10:1 are not uncommon, with all aberrations except distortion corrected to the necessary degree for 35 mm photography. As scientific instruments they suffer from varying amounts of distortion and must be calibrated *for each focal length* before being used for metrological work.

4.7 Practical aspects

Camera lenses come in all focal lengths from millimetres to metres. An astronomical telescope of 5 metres aperture with a Meinel–Courtès-type focal reducer attached becomes in effect a 35 mm camera with a 10,000 mm, F/2 lens. Since the early twentieth century the classical format for a photographic camera has been set at 24 mm × 36 mm, suitable for the Edison 35 mm film-width intended originally for motion pictures. Oskar Barnack (1872–1936), a microscope designer with Ernst Leitz of Wetzlar and an amateur photographer, gave the lead when he designed the Leica camera, and others followed. The format continued to the era of digital photography. Most camera lenses will focus objects well enough down to about 200 mm object distance. Macrophotography, as it is called, generally requires a dedicated lens so that object distances, usually less than the corresponding image distances, approach a few millimetres. This is on the borderline with microphotgraphy, which is the photography of a minute object, probably a specimen on a microscope slide.

[9] Thomas Rudolphus Dallmeyer (1859–1906), not to be confused with his father, John Henry Dallmeyer (1830–1883), also an optician.

A perfectly corrected lens would have all the ray paths from the object to the image the same optical path length. Lenses are not perfectly corrected, however, and a typical high quality 35 mm camera lens would have a variation of three or four wavelengths at any image point in the field, most of the variation being in the outer annulus. In practice this means that a point object at $-\infty$ would appear as an irregular spot with a diameter of about 3 μm. This is in addition to the diffraction spot, and the resultant spot-size is the convolution [10] of the two aberrations.

The diameter of the diffraction-limited spot is given by $d = 1.22F\lambda$ where F is the focal ratio and λ the wavelength. The spot diameter decreases with lower F-number whereas the geometrical aberrations tend to increase. Consequently there is an optimum F-number for high resolution photography. The factors determining this are

(1) the desire for sharpness in the image,
(2) the need to acquire an adequate exposure,
(3) the need to avoid camera-shake or blur if the subject is moving.

In general, for optimum sharpness of image, it is in the region of f/8 to f/11, which gives a spot-size of about 6 μm. For recording detail in a distant scene it is the *angular* resolution which matters and this is $\Delta\theta = 1.22\lambda/D$ where D is the lens aperture. The focal length can be made as long as necessary to reduce the geometrical aberrations to a trivial level. The diffraction spot diameter alone then determines the resolution. In the open air at sea level, atmospheric turbulence, even on a still, windless day, precludes a resolution finer than about 1 second of arc (1.7×10^{-6} radian), and for this reason an aperture of ~100 mm is quite adequate for long-distance photographic work.

In scientific photography a small pixel width is not necessarily an advantage, since with a properly designed system the data collection rate will probably depend not on the individual pixel size but on the whole étendue of the instrument, the *total* area of the detector multiplied by the solid angle subtended by the pupil at the centre of the detector. One must design an experiment carefully with this fact in mind.

4.7.1 Focusing the camera

If there is no autofocus (which only works if there are sharp edges in the frame), 'best' focus is by the marks on the index ring of the lens itself, or if necessary by eye, pure and simple; if necessary by the time-honoured method of

[10] See Appendix C for an explanation of convolutions.

inspecting with a hand-lens the image on a translucent focusing screen. Hand-focusing at infinity is the hardest, for example, when taking a picture of a cloudscape, for then focusing must be on a hard outlined object at a similar distance, with the autofocus switched off.

4.7.2 Depth of focus

This is *defined* as the error in plate (or CCD) position which will give an error of $\lambda/4$ in the wavefronts arriving at the focus. It is given by the expression:

$$\Delta f = 2\lambda F^2$$

where F is the focal ratio of the lens and λ is the wavelength of the light ($\sim$0.55 μm). It tends to be a doctrine of perfection. For a lens working at a focal ratio of f/2, the fepth of focus would be about $\pm$4 μm, approximately the diameter of a pixel, and a more practical figure would multiply this by a factor of 3 or 4.

4.8 Architectural photography

The photography of large buildings from a relatively close point of view presents some possible technical difficulties, regarding both distortion and depth of field. The traditional cameras to cope with these difficulties were the old 'tilt-and-shift' cameras and the 'rising-front' cameras, in which both the lens plane and the image plane could be tilted and moved. Dedicated plate cameras had these facilities and their application followed the *Scheimpflug principle*.[11] This relies on the geometrical fact that in the Gaussian approximation, the image plane, the lens plane and the object plane all meet on a line. This, plus the lateral shift of the lens, allows a tall building to be photographed – usually at a high focal ratio – without the usual 'converging parallels' caused by perspective.

4.9 Schlieren photography

This is a technique for visualizing small local changes in refractive index, such as are caused by the motion of air over an aerofoil section or the shock-waves produced by supersonic flow. The method is essentially that of the Foucault knife-edge test for focusing an astronomical telescope.

[11] Named after Theodor Scheimpflug (1865–1911) who applied it to the correction of distortion in military surveying applications.

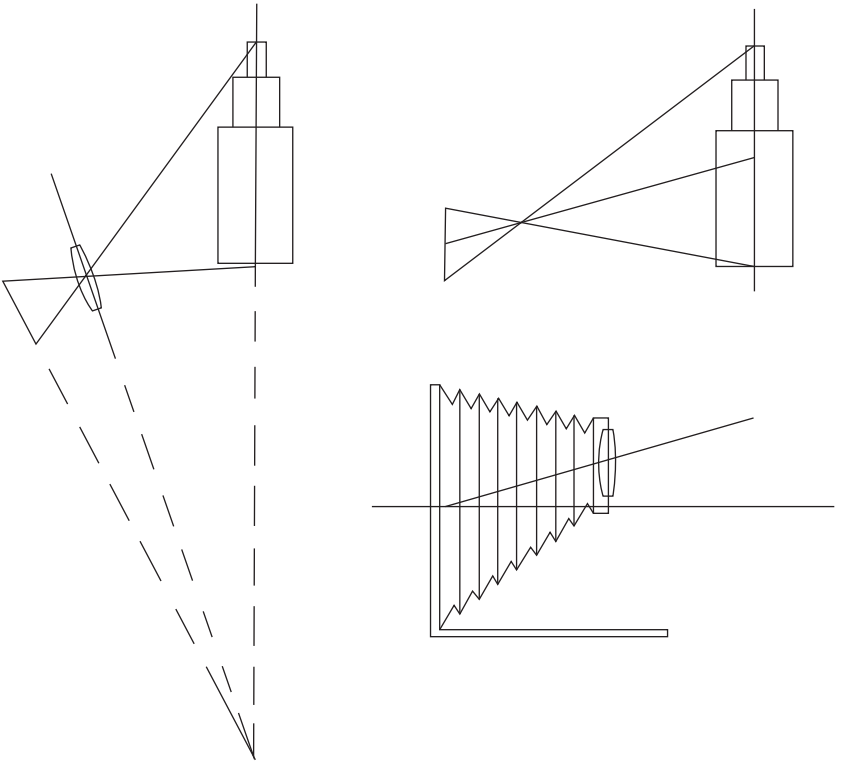

Figure 4.5 The Scheimpflug principle (left) and the rising front camera. This photographic trick relies on the fact that in Gaussian optics the object plane, the lens plane and the image plane, when they converge, meet along a line; and this line decides the geometry for keeping in focus the extreme points of a laterally extended object such as a tall building. In the limit the line is at $-\infty$ so that the planes are parallel and allow photography of tall buildings without converging perpendiculars, so long as the lens has no distortion. Large plate cameras in the Victorian era, with the Rapid Rectilinear lens, included the necessary adjustments for this kind of architectural photography.

An object in the form of a sharp edge, illuminated from behind with white light, is focused with high quality on to another edge, placed so that light is prevented from passing. Any disturbance to the optical path will then disturb the image and allow light to pass the second edge. A camera situated behind the second edge and focused on to the disturbed area will give a visible image of the cause of the disturbance.

The Foucault test similarly uses an edge to obstruct light from a star. The obstruction can only be complete when the edge is at the focus of the telescope

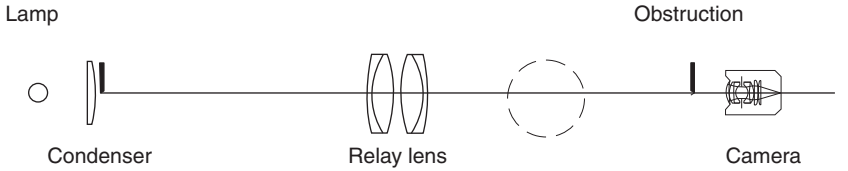

Figure 4.6 Schlieren photography basically means photographing an object and then replacing the developed negative so that no light is transmitted past it. Any disturbance to the air inside the optical system allows light to pass the edges of the negative, and if a camera is focused on to the disturbance, the disturbance itself can be photographed. In practice, a straight edge is used as object and another straight edge in the image plane traps all the light imaged by the system - unless there is a disturbance such as a flame, with the changing refractive index of the rising hot air.

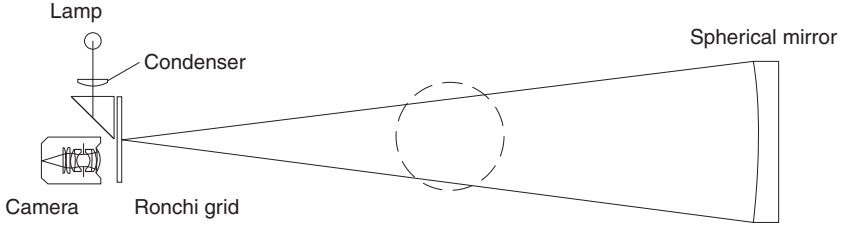

Figure 4.7 The Ronchi grid. This is a more elaborate version of the Schlieren method adapted to reflective optics. The grid itself comprises alternating opaque and transparent strips of equal width with a pitch typically of 0.5 mm. A spherical mirror, which can be several metres in diameter, forms a real image of the grid at its centre of curvature at a magnification of -1. The grid is adjusted so that one half exactly covers the image of the other half and traps all the reflected light. A large disturbing structure, such as a wind-tunnel, can be placed between the mirror and its focus, and changes in the refractive index of the air caused by rarefaction or compression can be seen and photographed.

objective, and the eye, looking at the objective, will see it uniformly illuminated by the starlight, gradually fading as the edge is moved across the image. Any defocus means that the starlight is 'wiped' aside by a dark curtain, moving across in one direction or the other according to the edge being inside or outside the true focus.

A *Ronchi grid*, a transparent slide of alternating clear and opaque lines of equal width, with a pitch of about 0.5 mm [12] can be used similarly.

[12] These grids can easily be home-made by old fashioned silver halide photography of a grid generated by a computer and printed on to A4 paper. Any fine monochrome emulsion is adequate.

4.10 The optical transfer functon

There are many ways of assessing the optical imaging quality of a lens. The quality depends not simply on whether it produces a sharp image. There are several factors involved and there is no unique definition. In the end it depends on whether the lens produces the results which the researcher desires.

The commonly adopted criterion is the optical transfer function, or OTF, and this derives from an earlier criterion proposed by Karl Strehl (1865–1940) for decribing the overall quality of a telescope mirror surface. It was intended as an alternative to a search for local surface imperfections by examining the surface with Fizeau interferometer fringes or a Foucault knife-edge test. A perfectly paraboloidal mirror would bring the image of a point object at $-\infty$ on the axis to a focus where diffraction was the only aberration, and the image would then be the 'Airy disc' of diffraction theory. The *Strehl ratio* is then defined as the peak intensity in an image point as a fraction of that of an Airy disc in the same place. It is a doctrine of perfection, however, as there is no easy way of measuring the ratio directly. A more practical method for assessing a camera lens quality is to measure its optical transfer function. This is done by examining the images of black and white grids of various degrees of 'fineness' or, to be more formal, of various spatial frequencies. If a screen is fine enough to be near the resolution limit of the lens it becomes impossible to distinguish between a grid on which the intensity varies sinusoidally and one made of alternate black and white stripes. The OTF is then a graph showing the image fringe visibility as a function of spatial frequency.

The *visibility* of sinusoidal fringes, v is defined as

$$v = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}$$

and a perfect fringe system would have a visibility of unity. Fringe visibility, incidentally, is a notion much used in interferometry, where it measures the *degree of coherence* between two interfering light sources – a concept to be discussed in Chapter 7. The image of a grid – indeed of any object – is the convolution of the original object with the image of a point object. The image of a point object, with all its aberrations, is called the 'point spread function' of the lens and it is this which would be an Airy disc if the lens were 'perfect'.

There is one peculiarity of the OTF worth mentioning: the curve of visibility against spatial frequency changes rapidly as a result of defocus, and as defocus increases, the OTF can become negative for a certain range of spatial frequencies. This results in 'false resolution' where a fine grid, on going out of focus,

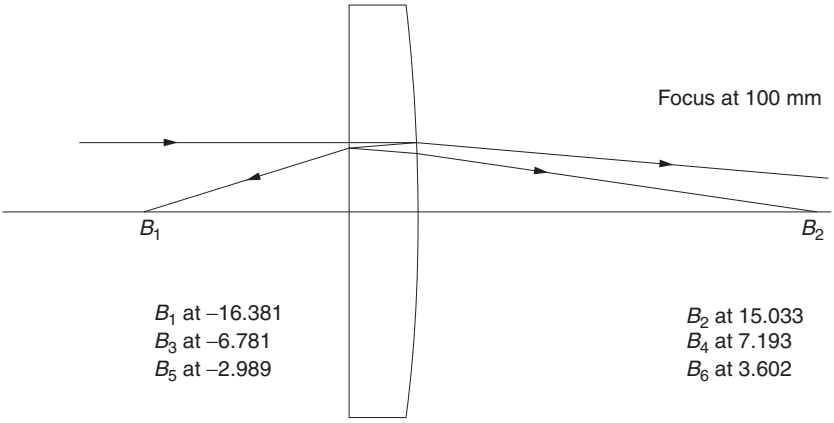

Figure 4.8 The Boys points of a lens. These are very faint and generally very close to the lens surface. They are best seen by imaging the Sun, and introducing a thin strip of paper into the system and searching along the optic axis.

comes back into focus – apparently – but with reversed phase: maxima where there were minima before and *vice versa*. A high powered telescope focused on to a newsprint monochrome photograph will show this effect.

The OTF has a particularly interesting application when a lens is being designed for a CCD camera. The designer need not struggle for an OTF which stretches beyond the spatial frequency of the CCD pixels but can design instead for a greater fringe visibility at lower spatial frequencies with quite a sharp fall at spatial frequencies above those which the detector can handle.

4.11 The Boys points of a lens

Boys points [13] form a topic somewhat neglected in optics books but nevertheless with interesting properties and uses in lens design.

Even with modern anti-reflection coatings there is a measurable amount of internal reflection at a transmitting surface. In other words, the curved surfaces of a single lens act as mirrors. Consider a simple biconvex lens. A collimated beam of light may be brought to a focus as usual, but a small fraction of the light will be reflected at the second surface and converge towards the first surface. There it is mostly transmitted to come to a focus on the input side. A small fraction of that beam, however, will be internally reflected at the first surface, transmitted through the second and come to a focus on the output side, the

[13] First described by Sir C. V. Boys (1855–1944) of the Royal School of Mines.

same side as the real image of the source. These two spurious images are the *Boys points* of the lens. The points are very faint but can be seen with a thin strip of paper held in the proper place and with a small point source. A zoom lens with many elements has a whole array of Boys points and quite frequently a video camera used in the open air on a fine day will find the Sun unintentionally in its field of view and the multiple internal reflections and the array of Boys points can be seen lying in a line across the image.

As a cautionary anecdote to would-be designers, the author, needing a particular image-size, once built a focal reducer to be attached to the 76-inch telescope at the Haute Provence observatory. It worked as designed, but superimposed on every astronomical image was a very faint ghost image of the primary mirror of the telescope, complete with the Cassegrain obstruction. It was a tribute to the extreme sensitivity of the LN_2-cooled CCD camera, and had been produced by scattering from the telescope mirror and various Boys point reflections in the reducer. It was pure bad luck that the image was in approximate focus on the CCD. A set of reference exposures, taken at random positions on the sky near the object of interest and median-filtered, provided the antidote, and by subtraction the ghost was exorcized (taking the sky background with it) on the final photometric frames.

5

The scientific CCD camera

5.1 The traditional silver halide camera

Silver halide photography, from its invention in the 1830s, relied almost exclusively on finely ground silver bromide crystals suspended in a gelatine emulsion. These, it was discovered, were altered by the absorption of light so that they could be reduced chemically to black, colloidal silver by a variety of sensitive reducing agents to produce a negative image in the emulsion. This could then be changed, usually by copying to positive images in transparent emulsion or on sensitized paper. The sensitivity was restricted at first to light of short wavelengths in the blue to ultra-violet region of the spectrum until a chemist, H. W. Vogel, at the University of Berlin, discovered the technique of dye-sensitization to make orthochromatic emulsions sensitive to yellow and orange light. Later improvements in this technique eventually made red-sensitive panchromatic emulsions available – at a price – in the early years of the twentieth century. S. M. Prokudin–Gorskii, [1] in particular, led the way by making simultaneous exposures through red, green and blue filters to give negatives, from which diapositives [2] could be made for simultaneous projection with three projectors and the same colour filters, on to a screen, thus producing the first colour photography slide shows.

So far as scientific photography is concerned, it was chiefly photometry which was of interest, and it was soon found that the density of the image was heavily dependent on various factors which had to be controlled carefully if consistent results was to be obtained. The chief factors were the choice of the particular reducing agent (including the pH of the developing solution)

[1] Sergei Mikhailovich Prokudin-Gorskii (1863–1944) studied chemistry under D. I. Mendeleev in St Petersburg.

[2] A transparent positive monochrome print from a negative.

and the time and temperature of the developing process. Reproducibility limits allowed a photometric accuracy of about 1% if strict care were taken. The further discovery that the number of photoelectrons released in a crystal to make it reducible to colloidal silver was largely independent of the size of the crystal, allowed emulsions to be made of different sensitivity and different spatial resolution, one being traded for the other. It was also found that the number of photoelectrons per unit area required to make the emulsion completely dense (saturated) was about 100 times the number which was needed to cause the faintest measurable change in density: in other words the dynamic range of the photometric process was about 100:1. The density, defined as minus the logarithm of the transmission, followed a curve, the Hurter–Driffield curve [3] which was more-or-less logarithmic in the middle portion. Such a curve was needed for every measurement and photometry was done by interpolating values between known levels of illumination.

5.1.1 Film speeds

Stemming from Hurter and Driffield's work, various systems such as the *Scheiner system* for denoting emulsion sensitivity arose. The Scheiner system was logarithmic and an addition of 3 to a film's Scheiner speed implied a doubling [4] of its sensitivity. The ASA (American Standards Association) system, on the other hand, was linear (and more popular) and evolved into the ISO (International Standards Organization) system which is still current. For example, a film speed of ISO 64 would require an exposure of 5 milliseconds at f/6.3 to record a landscape on a sunny day in a temperate climate. ISO film speeds went as high as 1200 and so-called 'surveillance' films could be 'forced' by harsh developers to ISO 2000, resulting usually in the fairly crude 'soot-and-whitewash' pictures of dimly illuminated night scenes.

Apart from a poor quantum efficiency and low dynamic range, photographic emulsions suffered from what is called 'reciprocity failure'. The number of photoelectrons required to form an image is a product of light intensity and time: the *exposure* is the product of the two. However, the concept fails at long exposure times and the 'speed' of the emulsion deteriorates rapidly at exposure times greater than a few tens of seconds. Various nostrums were applied to improve this, such as bathing the emulsion in an atmosphere of hydrogen for a

[3] Ferdinand Hurter (1844–1898) and Vero Charles Driffield (1848–1915) who brought some scientific sense to an analysis of the photographic process.

[4] The analogy with the decibel system in electronics is obvious.

few hours before exposure, but to no startling effect. In faint-light photometry, reciprocity failure always remained a great impediment.

Nevertheless, the ISO speed rating system is still used when comparing CCD cameras with their film-using ancestors.

5.1.2 Resolution

The spatial resolution of the emulsion is the factor which allows 'chemical' photography to continue. Crystals can be ground to fractions of a micrometer in diameter, and these archival emulsions were much used for data storage and microphotography.[5] One practical measure of the resolution of an archival emulsion was 'Bibles per square inch'. CCDs on the other hand are limited by the technology available for their manufacture and use, and the pixels are currently about 4–5 μm in width. This is becoming comparable to the 1–2 μm grain-size in 35 mm colour film.

Practical photography with the CCD is still advancing rapidly at the time of writing. Cameras for normal amateur photography, whether portrait or landscape, have CCDs approximating in size to half the 36 mm × 24 mm frame of the now traditional 35 mm camera, and with typically 16–20 megapixels. As the CCD is not colour sensitive it is covered with a 'Bayer mask'[6] which filters the light through red, green and blue filters in a rectangular pattern with the same size and pitch as the pixels. The consequence of this is that after processing in the camera's software, the spatial resolution is reduced by a factor of three, so that a fine rectangular grid is just resolved (and indistinguishable from a grid with sinusoidally varying brightness) with a pitch of 6 pixels. A 20 megapixel CCD will thus resolve horizontally about 600 line-pairs (of alternating black and white lines), and this is regarded as adequate for most *photographic* purposes.[7] Specialized cameras, for microphotography through a microscope for example, will use the whole possible resolution of the CCD chip by taking several photographs while moving the CCD through one pixel-width between exposures. The appropriate software does the necessary processing.

[5] Not to be confused with photomicrography – making large photographs of microscopic objects.

[6] Named after its inventor Bryce Bayer (1929–2012) of Eastman Kodak, the father of digital colour photography.

[7] For comparison, the author's old Leica 3C with a Taylor-Hobson 50.8 mm f/2 lens, resolved 120 line-pairs *per millimetre* on Kodachrome 25 in laboratory conditions. This is equivalent to a 12 megapixel CCD *without a Bayer mask* but in colour would require a 108 megapixel CCD to get the same resolution. The same camera/lens combination with Kodak Technical Pan film resolved more than 300 line-pairs/mm at f/2.

5.2 Scientific digital photography

For spectrography and photometry the CCD is used quite differently. There is no Bayer mask. After possibly several minutes' exposure, the charge of photoelectrons in each pixel is extracted, amplified and then sent to an analogue–digital converter to be recorded sequentially in a file. Generally the data are stored in FITS format [8] and there is ample software for inspecting, displaying and modifying the image and for examining and if necessary changing the contents of each pixel.

It is important to remember that the initial recording of the data is an analogue process and subject to the statistical variations which happen when analogue noise is added by the amplification process. This is in addition to the digital noise inherent in the detection of possibly very small numbers of photoelectrons in the individual pixels. The net consequence is that photometry, at low light levels especially, is no better than about 1%, much as in silver halide photography.

5.2.1 Cooled CCDs

Much of CCD photometry is done at very low light level, especially in astronomy. The signal-to-noise ratio is enormously improved by cooling the CCD chip, thereby reducing the accretion of thermal electrons and the statistical noise which goes with them. Two methods are used.

CCDs may be thermo-electrically (TE) cooled by the *Peltier effect*, [9] to temperatures between $-20°$C and $-75°$C which is cold enough for most scientific purposes. However, if ultimate performance is needed then a liquid nitrogen (LN_2) cooled refrigerator should be used, giving a temperature of $-196°$C. This allows the full quantum efficiency (of $\sim$0.95) of the CCD to be realized, which gives it a sensitivity equivalent to an ISO rating in the region of 4,000,000. This is a distinct advantage in very low light level astronomy for example. Stellar photometry was traditionally limited to stars brighter than $m_v = 27$ by the brightness of the sky background, since stellar images typically cover a circle of about 1 second of arc diameter and the sky background brightness, chiefly the night airglow, is typically one $m_v = 27$ star per arcsecond square. With CCD photometry, however, there is no question of reciprocity failure. The sky background can be subtracted (although leaving its noise behind) and stellar photometry at a magnitude as high as $m_v = 35$ is feasible. For example, with a 100-inch telescope the cycles of Cepheid variables

[8] **F**lexible **I**mage **T**ransport **S**ystem.
[9] Named after its discoverer J. C. A. Peltier (1785–1845).

in the M31 nebula in Andromeda can be measured and the distance of the nebula thereby accurately deduced. More immediately, a 15 minute photograph of a brick wall in a sub-basement room in complete darkness will show in outline the cement between the brickwork. The radiation comes probably from radioactive potassium in the cement, but if the room light has recently been switched off the phosphorescence of any of the materials of the wall may contribute. [10]

It is probably needless to add that such a cooled camera must hold its CCD in an evacuated dry chamber with a suitably heated double window and that such a camera housing is purchased and not constructed in-house. The technological art is quite subtle.

5.2.2 Long exposure

Exposure times above 30 minutes are apt to suffer noticeably from cosmic ray events (generally called CREs) and the usual trick is to take several photographs of, say, 15–20 minutes exposure and then to 'median filter' them, that is to take the *median* value [11] of the contents of corresponding pixels. For scientific work, where there is to be accurate photometry or mensuration, the data should be median filtered and there is on-line software available to do the filtering.

5.2.3 Manipulation of CCD frames

The FITS format is universal and there is much software available for modifying frames and for extracting experimental or observational data. In particular, successive frames may be added pixel-by-pixel to improve the signal-to-noise ratio, subtracted to show changes with time, divided by reference frames for correction of pixel-to-pixel sensitivity variations, median filtered, mean filtered, collapsed to a one-dimensional array for spectrography for example when the spectrum lines are parallel to pixel columns, and so on. It is quite feasible to write simple programs in a familiar computer language such as FORTRAN for other manipulations of the frame such as interpolation

[10] The author was astonished to find, by such subterranean experiments with his first LN_2-cooled CCD camera, that phosphoresence, usually at a very low level and with a half-life of a few minutes, is practically universal.

[11] The median, not the mean. If, for example, the same pixel in five successive exposures yields 121, 125, 118, 436, 127, the mean of these five is 185. It is evident that the fourth exposure has suffered from a CRE, and the median value of the series, 125, is chosen for the final photograph. Even in the absence of a CRE the median value is as good as the mean value for all practical purposes.

or cubic-spline fitting. The extraction of spectrometric data from a Fabry–Pérot étalon image of concentric circles to give a linear spectrum is a classic example of this. CREs which, in faint light photometry, give pixel contents far above the average, can be removed by requiring all pixels with contents above a certain number to be set to zero (which also incidentally removes all but the faintest stars from a photograph of the night sky).

5.3 Calibration

CCDs have a natural variation in individual pixel sensitivity and this must be taken into account when doing photometry. Astronomers are accustomed to taking what they call 'dome-flats', which implies exposing the CCD to the diffuse light of the telescope dome interior (with the lights on and the dome shutter closed to avoid acrimony and abuse from observers in adjacent domes). The idea is that each pixel receives the same level of light and that the comparative sensitivity of each pixel can be compared.

This in general is unsatisfactory. No camera illuminates its CCD uniformly. There is the '$\cos^4$ effect' on pixels at the edge of the frame. [12] There is the possibility of slight vignetting, especially with short focal length lenses at maximum aperture. Allowance for the pixel-to-pixel variation of sensitivity should strictly be done with the camera *with its lens in place* pointed at a featureless Lambert radiator. A distant matt white card is a rough approximation but it is difficult to illuminate it uniformly, and there are few natural Lambert radiators. A better approximation is a glass fish-tank filled with water holding a scattering suspension [13] and the camera lens held with its lens-hood close against the glass. The liquid is back-illuminated with a distant, small source of light and everything is to be done in a dark-room to eliminate untoward reflections.

Such a system compensates automatically for the geometry of the camera.

5.4 Radioactive lenses

One must be particularly careful not to use in a CCD camera, a lens which incorporates a thorium glass in any of its elements. Many lenses do. There may well be enough radioactivity to ruin a 30 minute exposure with a LN_2-cooled

[12] This is the geometric consequence of off-axis imaging of luminous objects, involving obliquity and the inverse-square law. It applies strictly to thin lenses and is only approximately true, especially with modern wide-angle lenses where the exit pupil varies with field angle.

[13] The author has found water with a bottle of milk mixed thoroughly to be adequate, but any colloidal suspension will do, provided it is dense enough to scatter light multiple times.

CCD by covering it with what are apparently CRE events. A Geiger counter is thus a useful accessory when choosing a lens. The author caused both concern and amusement when turning up at a famous photographic dealer's with a Geiger counter borrowed from the physics department Radiation Safety Officer, to choose an old-fashioned 35 mm F/2 camera lens free from radioactivity for his new LN_2-cooled CCD camera. A surprising number of lenses, some 50 years old, contained radioactive glass elements, probably thorium. This is not enough activity to be harmful – far below it – but it is enough to ruin a 30 minute exposure of the night sky and gegenschein.

6

Spectrometry

6.1 The prism spectroscope

The classical spectroscope was invented by Robert Bunsen (1824–1899) and Gustav Kirchhoff (1824–1887) working together at the University of Heidelberg. They used it to discover 'continuous' spectra, 'fluted' [1] spectra, line emission spectra and line absorption spectra. In effect they laid the foundations of spectroscopy as the backbone of scientific research. It is no exaggeration to say that the optical spectroscope has been the source of most of the data which have formed our understanding of the nature of the physical world.

The classical spectroscope comprised a slit, a collimator, a prism on a graduated rotating table (a *goniometer*), and a telescope with an eyepiece graticule. Its modern descendants are still to be seen in teaching laboratories where refractive indices are measured.

The quality of a spectrometer [2] depends on two things.

(1) Its ability to distinguish between two adjacent spectral lines, in other words its spectral *resolution*.
(2) Its ability to catch as much light as possible, i.e. its *étendue* or *luminosité* or light throughput.

A useful quantity in respect of the first quality is its *resolving power*, R, defined by:

$$R = \frac{\lambda}{\delta\lambda} \tag{6.1}$$

[1] Unresolved molecular spectra which at low resolution look like the flutings on a classical Greek column.

[2] The word *spectrometer* is used here indiscriminately to refer to spectroscopes, spectrographs and monochromators, all of which have seen service in the measurement of wavelengths and intensities of light.

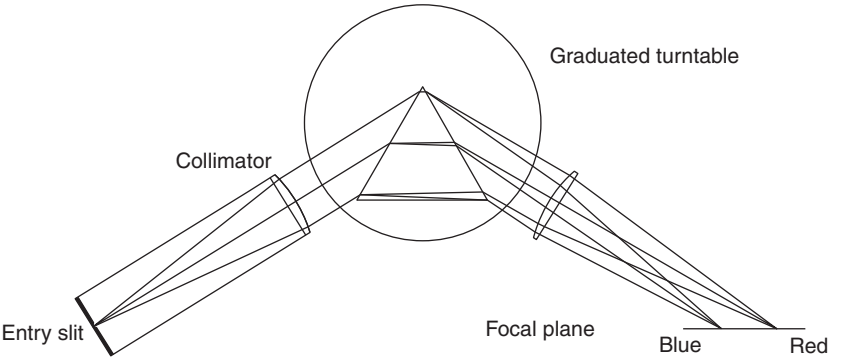

Figure 6.1 The traditional prism spectroscope. These were research instruments until about 1950, and may still be found in teaching laboratories. The basic sequence of entry-slit– collimator–disperser–telescope has never been challenged, although many different versions have been devised.

where $\delta\lambda$ is the difference between two wavelengths which are just resolved. Resolution here is defined by *Rayleigh's criterion* which says that two equal spectral lines are just resolved if there is a minimum of intensity between two maxima of 80% of those maxima.

Spectrometers generally can be categorized according to their resolving power. 'Low resolution' implies a resolving power of 1000 or less, able to identify the strong lines of familiar elements. 'Medium resolution' is between 1000 and 10,000, able to identify bands of diatomic molecules, and 'high resolution' implies anything over that, where isotopic and nuclear fine-structure and hyperfine-structure can be resolved and molecular electronic bands can be dissected to show rotational fine-structure.

These divisions are an arbitrary but simple guide to what may be required for a particular investigation.

Using Rayleigh's criterion, it is not difficult to show that for a prism spectrometer, irrespective of the prism angle, the resolving power available is given by

$$R = b\frac{dn}{d\lambda} \tag{6.2}$$

where b is the base length of the prism and $dn/d\lambda$ comes from the dispersion of the glass.

It is not difficult to show that for most types of spectrometer, irrespective of the optical design, the product of resolving power and étendue is a constant, a quantity known as its *efficiency* or simply as its RL-product. In the simple case

of the Bunsen–Kirchhoff spectroscope for example, if the slit width is doubled, then twice as much light gets through but the resolution is halved. Typical values for $dn/d\lambda$ can be deduced from typical figures for Schott SF57 glass:

$n = 1.83651$ at $\lambda = 6563$ Å (the C line in the Fraunhofer line list),
$n = 1.87205$ at $\lambda = 4861$ Å (the F line in the Fraunhofer line list),

so that a 1 cm base length prism of SF57 glass would give a resolving power of ∼2000 at 5700 Å, enough to resolve the D-lines of sodium at 5890 Å and 5896 Å. Oddly enough, the resolving power does not depend on the prism apex angle and various practical design matters collude to make the optimum apex angle ∼60°.

The choice of a spectrometer for a particular piece of research depends heavily on the type of data that is required. For simple detection of the presence of an atomic species, a resolution of about 0.1 nm (1 Å, 10^{-10} m) is sufficient. For the resolution of the lines of a rotation band of a diatomic molecule, a resolving power of $R \sim 50{,}000$ is required. To measure the temperature in a gas discharge plasma there is a well-known expression:

$$\frac{\Delta\lambda}{\lambda} = 7.16 \times 10^{-7}\sqrt{\frac{T}{M}}$$

relating the half-width [3] of a spectral line in angstroms with the Kelvin temperature T and the molecular weight M of the emitting species. For example, if the temperature of helium gas at about 10,000 K is to be measured by observing the red 7065 Å line, a resolving power of 250,000, or about ten times the FWHM, is the minimum required. In practice a resolving power of 500,000 is probably desirable. However, bearing in mind that resolving power is traded for étendue, the brightness of the source is a factor to be included when making the choice.

6.1.1 Improvements on the classical spectroscope

Given the resolving power expression above, there was a limit to the resolution, set by the size of glass prisms that could be made with sufficient optical homogeneity to allow light to pass without aberration. One way to overcome this limit was to use several prisms, and again it is easy to show that the resolving power is the sum of the various base lengths.

However, the limits on resolving power remained severe and Schuster's spectroscope gave only $R \sim 10{,}000$. The sodium D lines were resolved in his

[3] Strictly the FWHM, full width at half maximum.

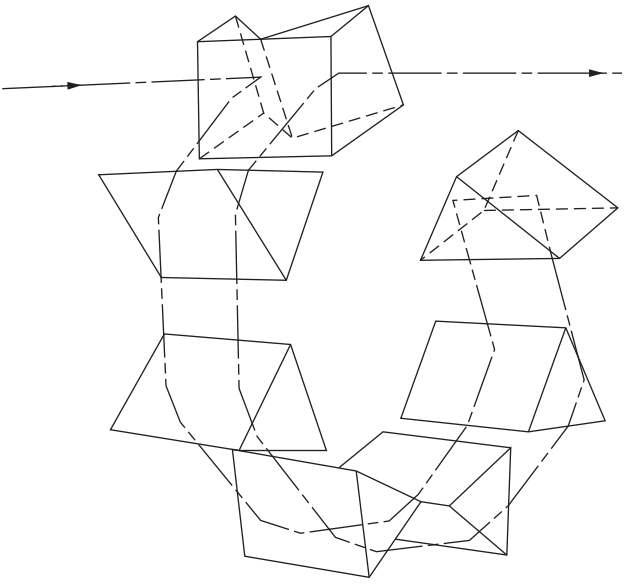

Figure 6.2 The multi-prism spectroscope. The resolving power is proportional to the base length of the prisms and is independent of the prism apex angle. Adding more prisms in this fashion allowed Sir Arthur Schuster, whose design it was, to increase the resolution of his late Victorian spectroscope tenfold. The final version had twelve such extra prisms, but without the benefit of anti-reflection coatings, it had severe curvature of the spectrum lines and abysmal transmission (<0.1%) and was never seriously used.

spectroscope, but after 26 refractions at 75% through a heavy lead-flint glass [4] the transmission was less than 0.1%. As the angle of incidence varied from the ends of the slit to the centre, spectral emission lines were also severely curved.

More seriously, the Littrow [5] configuration of a prism spectrograph halved the mass of glass required for the prism and gave resolving powers in the region of 20,000 in the blue and the near ultra-violet. Littrow quartz spectrographs were serious research tools until the 1960s.

The constant-deviation prism spectroscope is sometimes seen, with the Pellin–Broca prism as its disperser; but this too is now no more than a scientific curiosity, rarely seen except in occasional science museums or collections.

Few prism spectrometers remain in serious service and those are generally used for Raman spectrometry in teaching laboratories, where the absence of the 'ghosts' and side-lobes in emission-line spectra which can bedevil

[4] For one mode of polarization; the other mode is suppressed completely.
[5] The invention of J. J. von Littrow (1781–1840) of the University of Vienna.

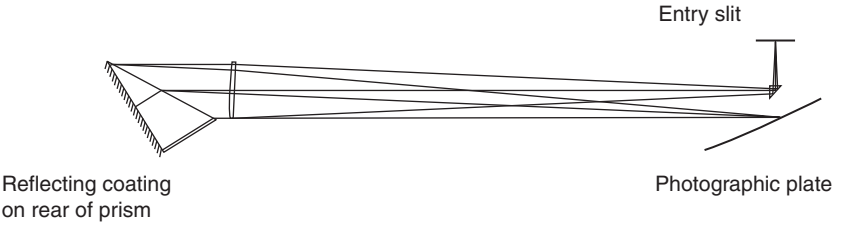

Figure 6.3 The Littrow prism spectrograph. To achieve the ultimate goal of diffraction-limited resolution, a high focal ratio was needed to reduce other aberrations to negligible levels. The Littrow spectrograph did this and at a focal length of 1.5 m gave a resolution of 0.1 Å in the blue region of the visible spectrum. The focal surface was slightly curved but 250 mm long photographic plates could be fitted without danger into the plate holder. This instrument was probably the ultimate performer among prism spectrographs.

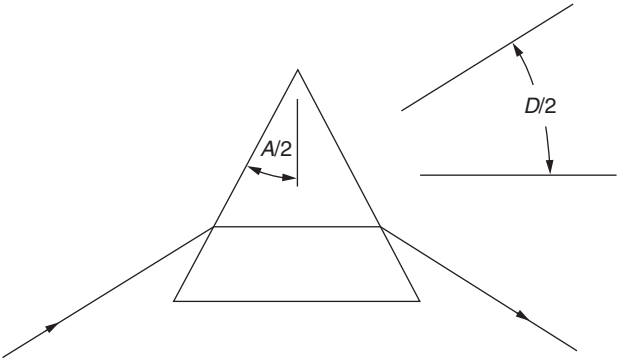

Figure 6.4 The geometry of prism dispersion theory. The symmetrical passage of light can be shown to give the maximum dispersion and at the same time minimum deviation. Elementary Euclidean geometry together with the law of sines leads immediately to the formula of equation (6.3).

grating spectrometry, is an advantage. Raman spectroscopy is now done by other means.

6.1.2 Visual measurements with a prism spectroscope

The standard expression for the refractive index of a prism is

$$n = \frac{\sin\left(\frac{A+D}{2}\right)}{\sin\left(\frac{A}{2}\right)} \tag{6.3}$$

where A is the prism apex angle and D is the angle of minimum deviation. The angles can be measured on the instrument usually to about ± 30 seconds of arc,

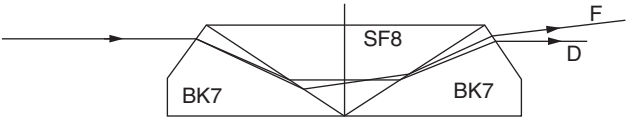

Figure 6.5 The direct vision spectroscope, with alternating prisms of low, high and low dispersion. The letters D and F refer to rays at $\lambda = 589.3$nm and $\lambda = 486.1$ nm, standard wavelengths derived from Fraunhofer lines in the solar spectrum. Likewise, BK refers to a low dispersion borosilicate crown glass and SF to 'short flint', a heavy, high dispersion glass. The device, due to Giovanni Amici (1786–1865), is an elegant solution to the problem of making a portable spectroscope for the inspection of sources of light and the identification of some elements, such as sodium, therein. A surprisingly high resolving power (∼500) is available from these little instruments.

so that refractive indices can be measured with an accuracy of ± 0.0001. It may happen that it is necessary or convenient to use a visual prism spectroscope to measure a refractive index or to measure an unknown wavelength. Once the prism table is levelled, the angles of minimum deviation can be found for each wavelength and noted on each of the two vernier scales on the goniometer. The entry slit is probably single-opening, that is to say only one jaw moves when the slit width is changed. All measurements should be made against the fixed jaw edge. There is no merit in using a narrow slit width for angle measurements, and the optimum technique is to have the eyepiece cross-hairs at 45° to the spectrum line, and to be sure to use the fixed slit edge throughout. This not trivial: the prism angle is measured by reflection from the faces so that there is a mirror image of the slit in view. The spectral line deviations are measured by transmission, and the fixed slit edge will be seen in the opposite sense to that seen earlier.

6.2 The direct vision spectroscope

This is generally an inexpensive hand-held device and may employ an Amici prism or, more likely, a replica transmission diffraction grating as its dispersing element. The retina is the detector. It has its place in the laboratory for the rapid identification of a light source if it is an emission spectrum from a laboratory lamp, or in a flame test for inorganic chemical analysis. It has its place too in the teaching laboratory as an introduction to spectroscopy. A properly made one will give a resolution of about 0.5 nm or 5 Å, enough to identify bright metal lines such as those from sodium, lithium, copper, zinc etc., in a flame.

The sodium D lines, being caused by a very strong resonance [6] transition, are almost always present, probably unresolved, at $\lambda = 589.3$ nm.

6.3 Diffraction grating spectrometry

Without going into detals of Fraunhofer diffraction theory, it is sufficient to say that if plane wavefronts of monochromatic light are incident normally on a grating the light will be diffracted from each facet of the grating in such a way as to produce plane wavefronts, but only at angles where the emission from adjacent facets is delayed by an integer number of wavelengths. The integer, usually denoted by m, may be positive, zero or negative and is the *order* of diffraction.

In practice the incidence angle is seldom normal to the grating and if the angles of incidence and diffraction are i and r with respect to the normal to the grating surface, the standard *grating equation* is

$$\sin i + \sin r = \frac{m\lambda}{a} \tag{6.4}$$

where m is usually 1 or 2 and occasionally more. Here λ is the wavelength of the incident light and a is the grating constant, the distance between adjacent rulings or facets. The angles i and r are assumed to be on the same side of the grating normal.

6.3.1 Plane gratings

Gratings are constructed in such a way as to give preference to one particular wavelength, directing most of the incident light into the first order of diffraction of that wavelength. This is called the *blaze* wavelength and the corresponding angle i is the *blaze angle*: but not only the first order of the blaze wavelength is diffracted. The second order of half that wavelength is diffracted in the same direction, and sometimes a filter, called an 'order-sorter' must be inserted in the optical train to suppress light of unwanted wavelengths in unwanted orders. Only one octave of the spectrum can be observed unambiguously. In the visible region, from 400 nm to 700 nm, there is less than one octave of wavelengths and the problem is not serious, but it is important nevertheless when discovering a previously unknown spectral line, to be sure that it is not a familiar line at half that wavelength, appearing in second order. Glass is mostly transparent to about

[6] An atomic energy-level transition from and back to the ground-state where no other return path is possible.

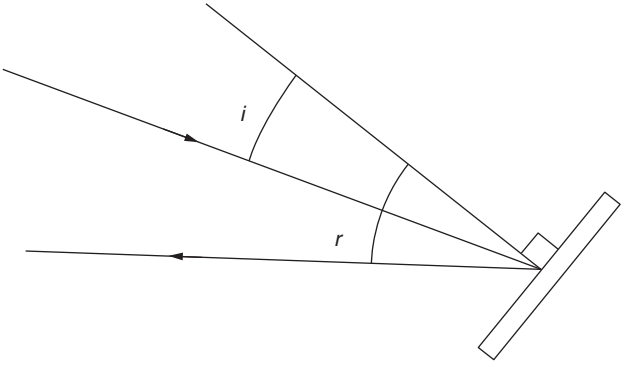

Figure 6.6 The incident and diffracted chief-rays on a plane grating, illustrating the angles of incidence and diffraction in the standard formula.

300 nm, and $\lambda = 350$ nm will appear in second order on top of $\lambda = 700$ nm in first order. The eye cannot see it, but the camera can. When complicated molecular spectra, especially that of molecular nitrogen, N_2, are being observed, utter confusion will reign unless you take care with the filtering or *order-sorting*.

Occasionally, to achieve a higher resolution, a spectrometer is worked in second or third order. Order-sorting then requires particularly strict attention, as a line may have, for example, a wavelength two-thirds of what is measured.

Gratings of research quality are almost always reflection gratings, the grooves being plane and inclined to the surface by the blaze angle. They are almost always *replicas*, made by copying in epoxy resin a master grating ruled on an optically flat ruled surface. Alternatively they may be made by making a hologram of two inclined wavefronts from a laser. The result then will be a grating automatically free from aberrations *at the laser wavelength* and with reduced aberrations in the neighbourhood of this wave length, always provided that the grating is installed in an optical system similar to that with which it was made.

Ruled gratings, on the other hand, require their aberrations to be corrected. This is done by using suitable mountings.

6.4 Grating mountings

The basic spectroscope layout perfected by Kirchhoff and Bunsen has never been surpassed. The sequence of slit–collimator–dispersing element–telescope is still the fundamental form of a spectrometer. The lenses of the telescope

have been replaced by mirrors, or possibly the same mirror passed twice, and the detector is now a CCD instead of a silver halide photographic plate.

At this point we must distinguish between spectrographs and spectrometers, otherwise known as *monochromators*. Both are in serious use, depending on the problem that is being researched.

Monochromators

These are spectrometers which select one wavelength and transmit its radiation through an exit slit – optically conjugate to the entry slit – to a single detector. A spectrum is then 'scanned', the grating being rotated slowly by a motor so that each wavelength in turn is presented at the exit slit for its intensity to be measured. Although this appears to be a wasteful way of obtaining spectra – nearly all the light is rejected instead of being recorded simultaneously as in a spectrograph – it has the advantage that focal ratio is not an important factor. The relative apertures of the optical components can then be made so large that aberrations are trivially small and the full resolving power of the grating can be realized. Monochromators intended for high resolution spectroscopy are therefore large instruments, typically with mirrors 3 metres in focal length and focal ratios in the region of F/20.

In simple diffraction theory, monochromatic light coming through a single parallel-sided slit is spread by Fraunhofer diffraction and when the light falls on a distant screen, the intensity follows the 'sinc-squared' pattern described in Appendix C:

$$I(\theta) = I(0).\mathrm{sinc}^2\left(\frac{\pi a \sin\theta}{\lambda}\right) \tag{6.5}$$

where θ is the angle of spread from the optic axis. Thus narrow slits give wide diffraction patterns and *vice versa*. In the case of a diffraction grating the 'slit' is the whole width of the grating and the very narrow diffraction pattern is the spectrum line of various wavelengths at various places on the focal plane. Being a sinc^2 function, it has side-lobes, or 'secondary maxima' on either side of the main line, and these have peak intensities 4% that of the parent. They are not likely to be seen in a CCD spectrometer, but an old-fashioned, finely tuned spectrograph will show them.

If, as a simple piece of geometry, the maximum of one sinc^2 function is superimposed on the first zero of a similar one, the combined intensity pattern will show a dip in the middle of about 20% of the maxima on either side. This is convenient for Rayleigh's criterion for the resolving power of a spectrometer

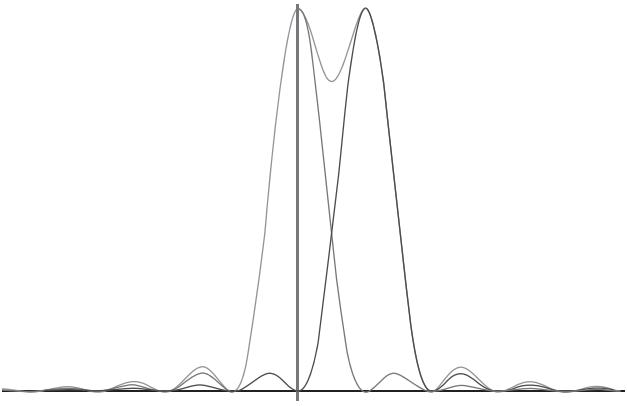

Figure 6.7 The shape of a monochromatic spectrum line emerging from a theoretically perfect grating spectrometer with an infinitesimal entry-slit width. The line width depends on the ruled width of the grating, and Rayleigh's definition of resolving power comes from the superposition of two such lines with the maximum of one lying at the first minimum of its neighbour. The sum of the two shows the 20% drop in intensity used by Rayleigh for his definition.

(which is why he chose it of course) as it allows the resolving power to be stated very briefly:

$$R = Nm$$

where m is the order of diffraction and N is the total number of lines on the ruled area. In a large grating of research quality the ruled width might be 200 mm and the ruling spacing, a, typically 833 μm, giving 1200 rulings/mm. The actual numbers[7] depend of course on the required performance. The theoretical resolving power is then $R = 120{,}000$ and at 500 nm (green light) one might expect to resolve two spectrum lines separated by 0.004 nm. In practice various imperfections conspire to reduce this resolution to about 0.005 nm but this is still adequate to resolve the hyperfine-structure of many atomic spectra. In the measurement of line widths caused by various effects such as temperature broadening, collision broadening, Zeeman effect or Stark effect, one should not rely on such a high resolving power, as the *shape* of the line is important as well as its half-width. Temperature broadening in particular can give a spectrum line a noticeably asymmetric shape and this is important as a diagnostic feature. To see the asymmetry you need the high resolving

[7] These numbers imply that all the rulings, placed end-to-end, would make a line 22 miles long: the length of the furrow gouged by the diamond ruling tip on the master.

power that a Fabry–Pérot étalon spectrograph or something similar can give, and these devices have their own chapter (Chapter 10).

Mountings for monochromators follow [8] the same general pattern as spectrographs. After a generation of designs based on the Ebert–Fastie layout the generally preferred configuration is the Čzerny–Turner arrangement, which allows an extra degree of freedom in design and is without the difficult complications of alignment which attend the setting-up of the Ebert–Fastie mounting.

Both arrangements allow for a 'flat field' although that is not of the highest importance in a monochromator. Both are governed in their behaviour by the grating equation (equation (6.4)), and from this equation you can see that for normal incidence ($i = 0$), there is a *maximum* wavelength given by $\lambda = a/m$, above which the grating will not diffract: for wavelengths longer than this it behaves as a plane mirror.

The Čzerny–Turner mounting has the advantage that by a suitable asymmetric arrangement of the components, meridional coma can be eliminated for one wavelength and substantially reduced throughout the range of the spectrum. This gives it about twice the resolution of Ebert–Fastie and other classic designs.

The flat field is essential for CCD spectrographs and if faint light sources are to be investigated, so is a low focal ratio. The focal ratio can be reduced without loss of resolution to $\sim F/2$ by a suitably designed focal reducer [9] and then the resolution available depends on the type of CCD that is used as a detector. Again there is a simple formula for the wavelength range, $\Delta\lambda$, arriving at an individual pixel:

$$\Delta\lambda = \frac{pa}{mWF_{cam}} \tag{6.6}$$

where p is the pixel width, a is the grating constant, m is the order of diffraction, W the ruled width of the grating and F_{cam} the focal ratio of the final camera lens. This expression is quite general, that is to say, it is independent of the particular grating mounting or general focal ratio through the instrument and only requires that the grating be the pupil of the system, which it nearly always is.

[8] Actually *preceded* the designs for spectrographs, since monochromators ruled the research scene from 1950 to 1990.

[9] J. F. James, *J. Mod. Opt.* **41**, 2033–2042 (1994).

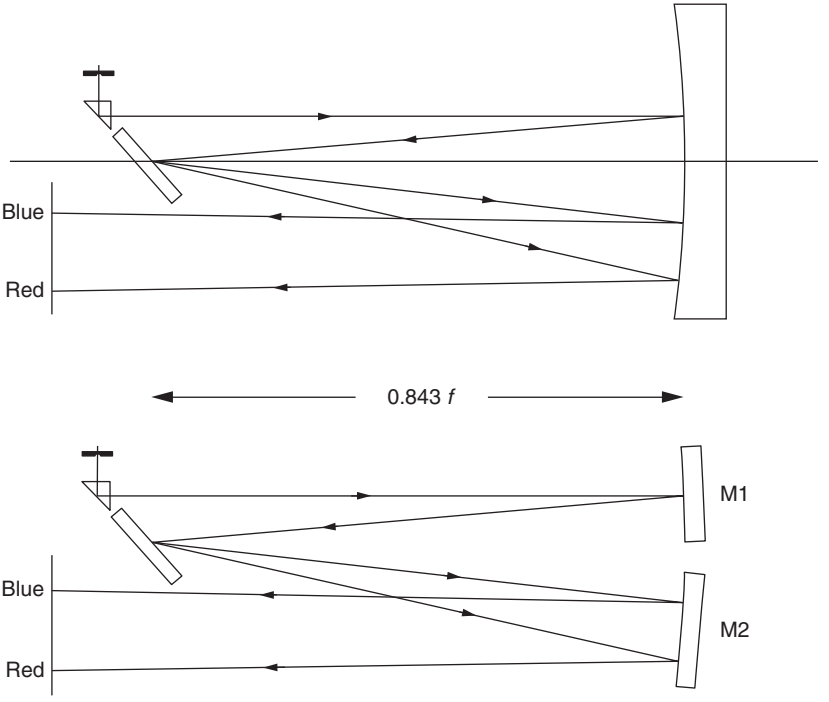

Figure 6.8 The Ebert (top) and the Čzerny–Turner configurations for a plane grating spectrograph. These are direct descendants of the traditional Kirchoff–Bunsen design in which the lenses have been replaced by concave mirrors and the prism by a grating; but it is essentially the same device. The Ebert was the first to appear, but was unaccountably lost until revived fifty years later by W. G. Fastie of Johns Hopkins University. Meanwhile, Čzerny and Turner had independently devised a basically similar two-mirror configuration which avoided or corrected the optical aberrations. In practice the single mirror of the Ebert design makes the optical path difficult to align and it has largely been replaced by the Čzerny–Turner version, where the two mirrors give an extra degree of freedom and allow better correction of coma. The Čzerny–Turner design is now almost universal in research surroundings.

6.4.1 Alternative mountings for diffraction gratings

The Čzerny–Turner mounting is practically universal, but the Littrow mounting will also be found, particularly in monochromators. It has the merit of being slightly more compact and of being practically free from geometrical aberration at the focus of the main mirror. This gives it a peculiar property of being useful as an *imaging* monochromator. An external rotating plane mirror can be made to scan an image of a bright, extended source such as the Sun across the top half of the entry slit. The lower half of the slit then transmits the

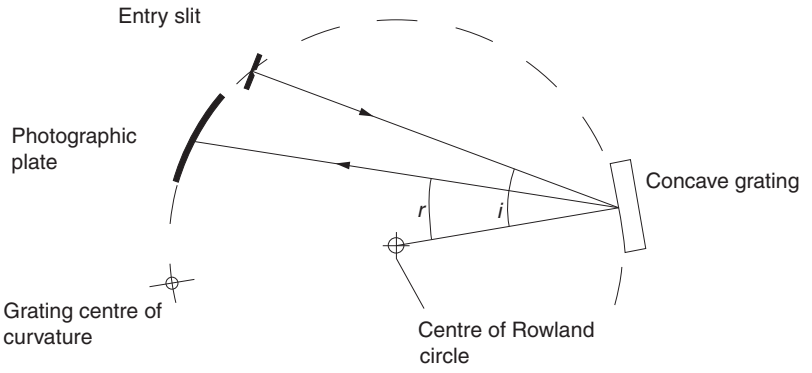

Figure 6.9 The concave grating spectrograph was devised by H. A. Rowland of Johns Hopkins University, who was the first to build a satisfactory grating-ruling engine. (It was remarked by the late H. J. J. Braddick of Manchester University, that designing a grating-ruling engine was the same problem as designing a screw-cutting lathe to be made of rubber.) The self-focusing property of the concave grating obviated the need for other optical elements in far ultra-violet spectrometry, where reflection and transmission coefficients were often very poor.

same slice of the object at a selected wavelength. This, scanned by the same rotating mirror, will give a monochromatic image of the object at a camera focused (through the rotating mirror) on to the slit.

6.5 Concave gratings

The concave grating was the invention of H. A. Rowland (1848–1901) of Johns Hopkins University. It is a diffraction grating ruled on to the surface of a concave spherical mirror, the rulings being equally spaced along a chord. He showed theoretically that the device needed no other focusing element, and that an entry slit placed on a certain circle (the Rowland circle) would give focused images of the slit at different wavelengths at points along the same circle. The Rowland circle is tangent to the grating at the grating vertex and its diameter is the radius of curvature of the grating. The usual grating equation, equation (6.4), governs the position of the spectrum lines along the circular focal surface.

Traditionally the grating was used with a flexible photographic plate or film fixed on the Rowland circle and there are various [10] geometrical arrangements

[10] J. F. James, and R. S. Sternberg, *The Design of Optical Spectrometers* (Chapman and Hall, London, 1969).

which allow it to be used as a monochromator. Its chief advantage is in vacuum spectroscopy in the extreme ultra-violet, at wavelengths between 1200 Å and 100 Å where there are no transparent materials except micrometre-thick collodion films, and best reflectivities are no more than ~10%. Typical grating ruled widths are 75 to 100 mm and radii of curvature are ~0.5 m to 1.0 m. In the historic past there have been heroic versions of the concave grating spectrograph with grating radii of 10 m, the spectrograph occupying a whole room with a Rowland circle 5 metres in diameter. A century ago this was the only way to get very high resolution (R > 100,000). These rooms have long since been converted, but traces of the original mountings may still be seen in older university buildings, for example in the form of unexplained brackets attached high on the walls.

6.5.1 High aperture spectrometers

These are chiefly for the study of faint extended sources such as the night sky, aurorae and nebulosities in astronomy. They may have a Schmidt camera to obtain a low focal ratio and there may be a transmission grating at the Schmidt stop – the pupil of the system. The focal surface is spherical, which makes it unsuitable for CCD detectors (quite apart from the heavy container for a LN_2-cooled CCD chip) and generally the performance of such spectrographs is below that of a Čzerny–Turner mounting with a LN_2-cooled CCD camera. If a medium resolution ($R \leq 10{,}000$) is sufficient it is not difficult to devise a simple mounting with a 50 mm F/2 camera lens to focus the spectrum on to a CCD. An example is illustrated in Figure 6.10.

6.6 Illumination of a spectrometer

As in all other photometric devices, the rule is that both the aperture (usually the entry pupil) and the field must be filled with light. In the case of a spectrometer, the aperture (and pupil) is the prism or grating face area *as seen through the collimator*, i.e. the projected area of the prism or grating on to the optic axis normal. The field is the entry slit, in this case long and thin. Typically the slit length is one-fiftieth the focal length of the collimator.

If the source is uniform and large compared with the entry slit, it may be placed close to it so as to fill the interior of the spectrometer with light. If the source is remote, then either the spectrometer is positioned to place the source on its optic axis, or a plane mirror redirects the light. Then a condenser lens is needed to image the source on to the entry slit. This condenser may not

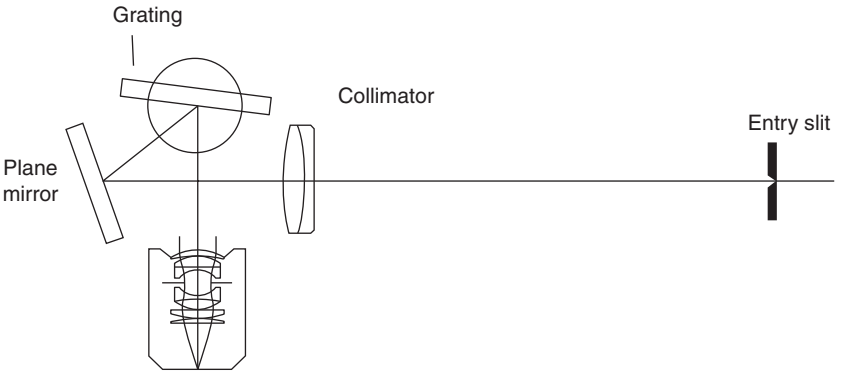

Figure 6.10 A simple high numerical aperture spectrograph for faint-object spectrometry. It uses a wider grating than comparable instruments and relies on the fact that the camera aperture, rather than the grating, is the defining stop of the system. With its long focal length collimator and standard camera lens, it is easy to design and construct without the need to explore its aberrations.

have enough aperture to allow light to reach every part of the grating or prism aperture. Another lens is then needed to image the condenser on to the grating. This second lens, called a *Fabry* lens after its inventor, is generally placed near the entry slit, usually just in front of it. It acts in the same way as the field lens of a telescope eyepiece and forms an image of the condenser big enough to fill the grating aperture. Bear in mind at this point that the *minimum* distance between an object and its image cannot be less than $4\times$ the focal length of the lens. The Fabry lens must be chosen accordingly.

Sometimes a source under investigation may be imaged on to the entry slit by the condenser at a magnification sufficient to fill the length of the slit with its image. If the source is not uniform – a Bunsen burner flame for example – the spectrum will show these variations. The blue cone at the base of the flame will show the band spectra of the molecules in the combustion, the blue *Swann bands* of C_2 for example, and the violet bands of the hydrocarbon radical CH. At the yellow top of the flame the hot soot particles will show the black body spectrum appropriate to their temperature.

It may well be that the condenser lens which first images the source on to the slit is big enough to fill the grating with light, in which case a Fabry lens is superfluous. If a distant extended source is being observed, such as the night sky for example, there need be no condenser, but there should be a lens to image the sky on to the entry slit. This is to ensure that all parts of the grating are looking at the same place on the sky. In a spectrograph such as the one

shown in Figure 6.10, where the camera iris and not the grating is the iris of the system, different parts of the grating will receive light from different parts of the sky unless this lens is in place.

The question arises as to the correct slit width for maximum resolution. A short answer in the days of silver halide photography was that it should be such that light from a monochromatic source would be diffracted through the entry slit by such an angle that the first maximum of the diffraction pattern just filled the grating ruled width. In a CCD spectrograph, however, the final image of the entry slit should have the same width as one pixel. A narrower slit will transmit less light and will not improve the resolution. It is a doctrine of perfection, however, as the slit image will not have the same width at every point on the spectrum. If the entry slit has an angular width δi its image will have an angular width given by differentiating equation (6.6) to give:

$$\delta r = (-\cos i / \cos r)\delta i \tag{6.7}$$

and this changes as r changes across the width of the CCD. If quantitative photometry is the aim, this is an important factor and not to be neglected. However, reference spectra with a known standard light source are always the best means of comparison if double-beam spectroscopy is not available.

In practice the full theoretical resolution is rarely achieved, but careful adjustment of the instrument will yield about 80%. Spectrographs are not often used at the limit of their theoretical resolving power.

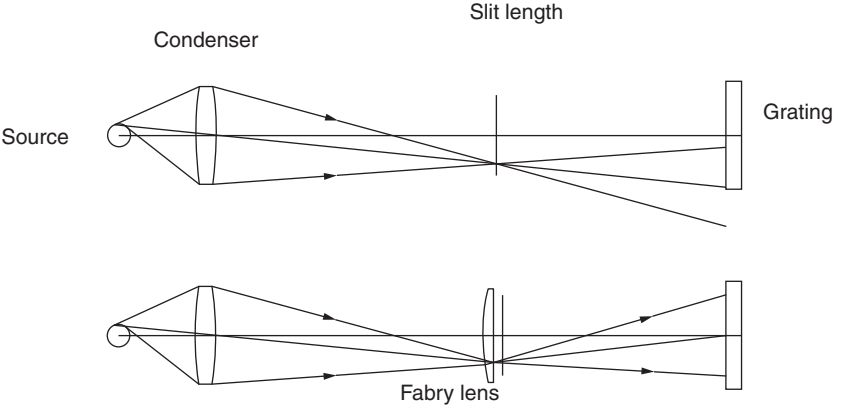

Figure 6.11 The Fabry lens. This is a neglected but important item in improvised laboratory spectrometry. It ensures that light which has passed the condenser and the entry slit will arrive at the grating aperture, thus filling the instrument with light.

6.6.1 Fibre-optic coupling

A fibre-optic input to the entry slit may be an advantage. Some spectrographs are already adapted to this and the entry slit can be replaced by a slot for the ferrule at the end of the optic cable. However, it is well to bear in mind that if there is to be high resolution, the input slit will be narrow, its width measurable in micrometres or at most a few tens of micrometres, and the end of a fibre bundle emits its light into a cone of about 10° radius, or about f/3 in camera terminology. A lens can be used to change the focal ratio to match the M1 mirror focal ratio (otherwise most of the emitted light misses the mirror) but at the expense of a large image of the bundle end, [11] requiring a wider slit, which in turn may compromise the spectrograph resolution.

6.7 Grating ghosts

Single strong emission lines are sometimes accompanied in diffraction-grating spectrographs [12] by faint images called 'satellites' or 'ghosts', lines produced by the same parent line, close on either side of it. There are several possible causes for this.

(1) They may constitute a genuine spectral feature such as electronic or isotopic hyperfine-structure in the line.
(2) They may be secondary maxima of the diffraction pattern produced by the instrument itself.
(3) They may be the result of a periodic error in the screw of the original grating-ruling engine. This is the optical analogue of frequency modulation in radio technology, in which the periodic change of frequency produces 'side-bands' to the carrier frequency. In spectroscopy, these satellite lines are known as 'Rowland ghosts', after the same H. A. Rowland who invented the concave grating and who incidentally built the first of the modern, precisely engineered grating-ruling engines. These ghosts can wreak havoc with old-time Raman spectroscopy, where the detection of faint satellite lines is the be-all and end-all of the investigation. Raman spectroscopy is no longer done this way, but anomalous fine-structure from this cause must be allowed for in the ordinary course of spectroscopic work.

[11] The consequence of the law of conservation of étendue.

[12] These are absent in prism spectrographs but secondary diffraction maxima may be seen in long exposures because of diffraction by the whole prism aperture – the dimension which determines the limit of resolution in the first place.

(4) They may be the result of a semi-periodic error in the ruling engine such that alternate rulings are of slightly different width. They are lines which appear at *half* their proper wavelength, as if there were an order $m = 0.5$. They are called 'Lyman ghosts' and are rare [13] to the point of being legendary.

6.8 The alignment of a grating spectrograph

Large spectrographs will have been aligned optically when first installed and generally will not need re-alignment unless they are very old or have been moved or dropped. Assuming that the manufacturer's handbook has been lost, the alignment can be checked without making any adjustment to the instrument and if it is of the Čzerny–Turner configuration it can be re-aligned if necessary without difficulty.

6.8.1 Checking the optical alignment

If the spectra are out of position or 'soft' in focus, there may be a misalignment.

In a *monochromator* the test for focus is simple, and is in essence the Foucault test for finding the true focus of an astronomical telescope. [14] In this present case the spectrometer entry slit takes the place of a star.

Put a bright, emission-line light source, a *low pressure* Hg or Na lamp for example, at the entry slit (which can be opened wide), close enough that the whole instrument is filled with light. Open the exit slit wide and turn the wavelength dial until the bright line appears there. Check that it is where it should be, i.e. centred on the exit slit and not too high or too low. Then look *into* the instrument *through* the exit slit *at* the M2 mirror (or at the Ebert mirror if it is of that mounting) and turn the wavelength dial slowly. The M2 mirror should appear full of light when the line is presented, but as the line moves away with the tuning of the wavelength, if the spectrometer is well aligned the light should vanish uniformly on the mirror, perhaps showing 'zones' because of spherical aberration. On the other hand, if a black curtain appears to sweep sideways across the mirror, then there is a defocus. Which way it sweeps depends on the number of reflections which have happened inside the instrument, and a refocus-and-test will find the true focus. There should be adjustments on one

[13] Although the author recorded them very faintly while working with a concave grating spectrograph in Ditchburn's Vacuum UV Laboratory in Reading University during the last century.

[14] French astronomers have a verb (not in the dictionaries) 'foucaulter', to focus a telescope using this test.

or both slit mountings for this. If necessary, turn the slit mountings about the local optic axes to make the entry slit and its image parallel. Rarely, the slits are fixed in position and then the whole instrument must be checked for misalignment and the M2 mirror moved.

In a *spectrograph* a misalignment will show as an incorrect position of the spectrum on the photograph (silver halide or CCD) and/or a softening of sharp emission lines from a low-pressure test lamp. If a minor adjustment of the M2 mirror or the camera focus does not correct it, it is probably as well to re-align the whole optics from the start.

6.8.2 *Ab initio* alignment of a spectrograph

The basic idea is to ensure that each optical element receives light at the centre of its aperture or field stop, and directs the light to the centre of the next aperture or field.

A HeNe laser (not a solid state pointer-laser) will give a practically monochromatic spectrum line and acts as the optic axis.

(1) The laser beam, shone through the middle of the entry slit and directed at the centre of the M1 mirror, should be reflected to appear at the centre of the grating ruled area. If it does not, the M1 mirror mounting screws need adjusting until it does. The accuracy of adjustment needed is not great: ± 1 mm is sufficient in positioning the laser spot. Cardboard cut-outs to fit over the mirror or grating, with centres marked as targets, will help to aim the laser beam.
(2) The grating is turned on its mounting until the diffracted laser beam appears at the M2 mirror, and again the beam should arrive at the vertical centre as it sweeps across the mirror with wavelength tuning. If it does not, the grating mounting tilt (not yaw) must be adjusted.
(3) The beam reflected from the M2 mirror centre should appear at the proper position on the photograph. In old-time spectrographs a ground-glass screen in the focal postion will help; otherwise photographs of the field in the spectrograph CCD camera should show the laser image at its proper position on the spectrum. If it does not, the M2 mirror mounting should be adjusted.

Note: things must be done in this order to get a proper alignment.

It is highly unlikely that misalignment would be accompanied by a defocus. The focus of the M1 mirror can be checked very accurately provided the grating can be turned so that the light is reflected back to the entry slit. Then a *Ronchi grid* (q.v.) is placed at the entry slit position, and light shone through

it (via a small 45°−45°−90° prism) at the M1 mirror. The light is reflected back from the grating at magnification −1 and if the focus is perfect the grid will be filled with light. If there is a focus defect the grid will show alternating stripes – a Moiré effect. The process is equivalent to Foucault's method for finding the focus of a telescope. It will only work for the M1 mirror, however.

6.8.3 Thermal insulation

This is sometimes neglected but is important at high resolution when wavelength measurement is crucial. The spectrograph must be at a uniform temperature. Sunlight coming through a window, and in the case of a high resolution spectrograph even the heat of a soldering iron carelessly left on nearby, or the body-heat of a worker standing close by, can warm one side of the instrument enough to change the wavelength calibration. If an experiment is to be highly repeatable, then a thermostat system for the interior is worth considering. A well-stabilized spectrograph with a resolving power of 20,000, illuminated by daylight, can detect the Doppler movement of a solar Fraunhofer line as its wavelength changes by $\sim 6 \times 10^{-3}$ Å from morning to evening.[15] The image of the Fraunhofer line will move by far less than a pixel width, but there is enough redistribution of the light between adjacent pixels to see the effect, provided it is done carefully, with deep Fraunhofer lines,[16] and with the spectrograph held at a constant temperature. Subtraction of the morning and evening CCD spectra thus obtained can be remarkably revealing.

[15] At 53° latitude.
[16] With the Hβ 4861 Å line of hydrogen for example.

7

Interferometers and their uses

7.1 Interferometry

Interferometers, as the name implies, rely on the wave nature of light and of the constructive and destructive interference between wave-trains. They are devices, often of great ingenuity, with many applications to fine measurement ($\sim 10^{-9}$ m) or ultra-fine measurement ($\sim 10^{-12}$ m) and as such they have led the way to many advances in science, not least in provoking the theory of relativity and in confirming the predictions of quantum mechanics. After overcoming formidable technical problems in construction and stability, interferometers are now in everyday use, particularly in spectroscopy, fine measurement and quality control.

7.2 Oscillation, phase and phase-difference

The word 'oscillator' applies as much to the electric vector of a beam of radiation as it does to a swinging pendulum. Two oscillators are 'in phase' when they reach a maximum displacement in the same direction at the same time and are 'out of phase' if their displacements are exactly opposite. We then say that their phase-difference is 180° or that they have π radians phase-difference. Their *phase-difference* is measured as an angle because it is convenient to use the mathematics of a circle to do so: there are no physical lines or angles involved. It is also convenient to imagine a vector in Cartesian coordinates, with one end fixed at the origin, rotating at the frequency of the light beam ($\sim 10^{14}$ Hz) and to take the x-component of the vector as the amplitude. Two beams of light arriving at a point can then be represented by two such vectors with a fixed angle between them, and rotating together. The fixed angle is the 'phase-difference' and the resultant amplitude is the vector sum of the two.

The *intensity* of the light is proportional to the *square* of the amplitude of the electric vector.[1] When two beams arrive at a point and are in phase, the intensity of the light is four times that of either beam on its own.

7.3 Coherence

Two wave-trains *interfere* with each other if crests from one train arrive at a place simultaneously with troughs from the other. The two amplitudes are opposed and cancel out and there is no net electric field at that point. For this to happen *at all times* the two wave-trains must be *coherent*.

No real wave-train is purely monochromatic: there are small, random changes of phase from moment to moment and from point to point. Two wave-trains are said to be coherent if the small changes in one are exactly matched by the changes in the other. Going back for a moment to the rotating vectors model, the two rotating vectors are locked together with a constant phase-difference, no matter how their actual rotation speed may change slightly moment by moment.

We can also speak of 'partial coherence' where the angle between the two vectors is not rigidly fixed but varies randomly by a small amount (small compared with 360°, that is) so that the resultant amplitude changes length slightly from moment to moment and averages out when photographed or otherwise observed. The consequence is that the interference fringes are less distinct. We measure the degree of coherence by the *visibility* of the fringes, defined by

$$V = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \tag{7.1}$$

and obviously $V = 1$ for perfect coherence, and $V = 0$ for two completely incoherent sources.

7.3.1 Coherence length and coherence area

If two different portions of the same wavefront are compared, the degree of coherence is found to decrease gradually as the separation of the two points increases and we speak of the *coherence area* as the size of an area over which measurable fringes can be seen. This area is not fixed on the wavefront, and two different interferometers looking simultaneously at the same source – a star for instance – will measure the same coherence area. For example, light from the star Betelgeuse, observed by Michelson, showed a 'coherence area'

[1] Again this is true of all oscillators: the bob of a pendulum with twice the amplitude of swing has four times the kinetic energy as it passes the equilibrium point.

of several metres diameter, from which he deduced that the angular diameter of the star was about [2] 0.05 seconds of arc.

Coherence length is measured by the fringe visibility when one portion of a wave-train is compared with a later portion, as is done in an interferometer when there is a path-difference. Monochromatic fringes from an atomic transition lose their coherence when the path-difference is about 3 metres, implying that the 'length' of a photon is about 3 metres, or, more plausibly, that the transition time of the emission [3] is about 10^{-8} s.

Commonly, coherence is achieved experimentally either by *division of wavefront* or by *division of amplitude*. Interferometry is based on one of these two methods.

7.4 Division of wavefront

Interference can be obtained by intercepting two separate areas of a plane wavefront and then superimposing one on the other by some optical device. Early experiments on interference by Thomas Young (1773–1829) involved shining monochromatic light from a point source through two adjacent slits. Each slit diffracted the light into a spreading beam and where the two beams overlapped, fringes appeared, light or dark according to the path difference of the light from each slit at a point on the screen.

Augustin Fresnel (1788–1827) and Humphrey Lloyd (1800–1880) relied on the division of wavefront by forming virtual images of a single monochromatic light source, and superimposing the light from each to form interference fringes. Fresnel used a prism with an apex angle close to 180° and Lloyd used a mirror at grazing incidence. In the two latter methods one part of a wavefront was diverted and made to overlap with another part while tilted at a small angle. Interference and reinforcement of the amplitudes could then be seen on a suitably placed screen.

Several useful devices were invented based on the division of wavefront. First among these was the Rayleigh refractometer, a variation on Thomas Young's two-slit experiment, which proved useful in measuring the refractive index of a gas as a function of pressure. It had a cell in one arm which could be filled with a gas at increasing pressure with a consequent increase of optical path and so a retardation of phase.

This device should not be confused with other refractometers such as the Abbé refractometer, which measures the refractive index of a liquid by

[2] The diameter of Betelgeuse is actually variable.
[3] Also known as the lifetime of the excited state.

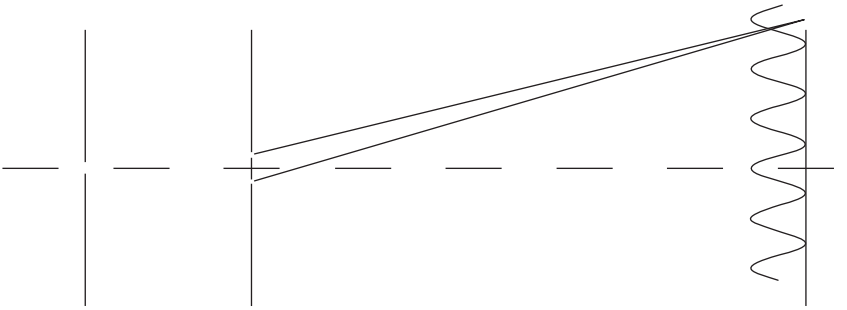

Figure 7.1 Thomas Young's arrangement for observing two-slit interference. The narrow slit on the left is illuminated with monochromatic light. The broad diffracted spread of light after passing this slit illuminates two further slits with wavefronts which are coherent. The further diffraction pattern coming from these two slits is crossed with interference bands. If the path-difference is an integer number of wavelengths, the amplitudes add and the intensity is four times that of single-slit illumination. On the other hand, if the path-difference is an integer plus half a wavelength, the amplitudes oppose and cancel, giving darkness at that point. The same interference effect can be obtained by various devices, for example, by reflecting the light from a narrow slit at glancing angle in a plane mirror, thus providing two closely spaced coherent sources, one virtual, which will interfere in the same way. This was Lloyd's mirror experiment. Fresnel used a prism with only a slight dihedral angle between two faces to produce two closely spaced virtual images of one source. The effect was the same.

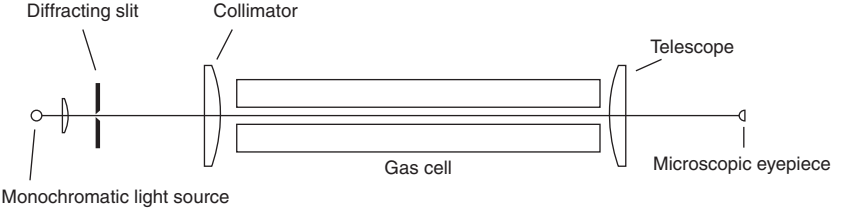

Figure 7.2 The Rayleigh refractometer. The collimating lens sends plane, monochromatic, coherent wavefronts through two cells, one containing a gas at variable pressure, the other containing air at room temperature and pressure, or possibly a vacuum. The focusing lens provides an image of the diffracting slit which is crossed with interference fringes. The image is minute and the fringes closely spaced, so that they must be observed through a magnifier.

observing the critical angle for total internal reflection at the hypoteneuse of a right-angle prism. The Rayleigh refractometer, like its successor the Jamin refractometer, is a much more sensitive instrument, used for gas analysis.

Following the discovery of a method of chemical deposition of thin silver films on to glass, division of amplitude has proved a much more powerful

interferometric technique. Division of wavefront lives on, however, in such instruments as the Rayleigh refractometer and the Michelson stellar interferometer for measuring the angular diameters of stars.

7.5 Division of amplitude

This is achieved by sending a monochromatic beam of light through a *beam-splitter*, a transparent surface treated with a partially reflecting coating which reflects half the incident power and transmits the other half. The two beams thus created are coherent and if they are recombined later and focused on to a screen after following approximately equal optical paths, they will show *interference fringes* which can be displayed on a screen or observed by eye or camera. There is a myriad of ways of doing this, each with its own purpose. Modern beam-splitters are made not by silver-film deposition but by vacuum deposition of suitable materials, and high precision and low absorption are both routinely achieved.

7.6 Interferometers

You are likely to meet four types of interferometer in the laboratory:

(1) the Fizeau interferometer for checking the flatness of an optical surface,
(2) the Michelson interferometer, a two-beam device for measuring wavelengths and for *Fourier spectrometry*,
(3) the Mach–Zehnder interferometer, a two-beam device for measuring small changes of optical path in one arm, often with fibre-optic cables as the two light-paths,
(4) the Fabry–Pérot étalon, a multiple-beam device for very high resolution spectrometry, high speed, time-resolved spectrometry and the spectrometry of very feeble light sources.

There are many other types of interferometer of course, each with its own special application. The Twyman–Green interferometer is a variation on the Michelson configuration, and is used for checking optical surfaces and the wavefront shapes they produce. The Mach–Zehnder configuration, like the Michelson, can be used for Fourier spectrometry, chiefly in the far infra-red. The Sagnac or ring interferometer is highly sensitive to rotation and is used in inertial navigation systems; and there are various configurations using a laser for making holograms and in particular for making holographic diffraction gratings.

7.6.1 General interferometer theory

When two beams of monochromatic light coming from a common source are divided and then recombined after passing through two unequal optical paths, the two amplitudes will be in phase and will add to double the amplitude to the electric vector, provided:

$$2nd\cos\theta = m\lambda \tag{7.2}$$

where d is the *axial path-difference* along the paths traced out by the two beams while they are separated, n is the refractive index of the medium through which they pass, θ is the angle between the wavefront normal and the optic axis, λ is the wavelength of the light and m is called the *order* of interference, the number of wavelengths of path-difference. For constructive interference resulting in a maximum of intensity, m is an integer. In practice and depending on the type of interferometer, m, the order, can be any integer from zero to $\pm 5{,}000{,}000$ or more.

Equation (7.2) is the fundamental equation of interferometry and applies to all interferometers, regardless of type.[4] It is based on the useful mnemonic: 'phase-change equals $2\pi/\lambda\times$ path-change in any beam of monochromatic light'.

There are five possible variables in the equation. In any interferometer three are held constant, one is an independent variable and the interferometer measures the fifth as a function of the other. For example when a Michelson interferometer is used as a metrological device $\theta = 0$, $n = 1$ (measurement is in a vacuum), $\lambda = 605.78$ nm (an international standard wavelength of ^{86}Kr) and the order m is counted electronically as d, the path-difference, is changed. The instrument, in other words, measures the number of wavelengths in a distance d. As another example, in the Fabry–Pérot étalon if d, m and n are held constant, λ is measured as θ varies: the device is a spectrograph, plotting wavelength, λ, radially outwards[5] as a function of θ. In the early days of interference spectroscopy the étalon was put into a pressure chamber. Then $\theta = 0$, d and m are fixed and λ was plotted as a function of n as the pressure was changed in the chamber.

[4] Purists will say that it does not apply to the *spherique*, a device invented by the great eccentric spectroscopist, Derek Jackson. His recreations included steeplechasing (he rode in the Grand National three times), antique silver and marriage (he was married six times, often to some very famous ladies). He was also famous for his brusque impatience as session chairman in optics conferences.

[5] Caution: the wavelength *decreases* as θ increases

7.7 The Fizeau interferometer

This is normally a common-user test instrument in an optical workshop, enabling the flatness of a glass surface to be assessed against a standard optical flat. Light from a monochromatic source, traditionally a mercury lamp with a green filter to isolate the bright $\lambda = 5461$ Å emission line, but now usually a diode laser, is collimated by a figured lens to eliminate spherical aberration, and the light is reflected from a silver-coated reference-flat and from the surface under investigation held above it. The gap between the two surfaces is about 1 mm, is nearly constant and adjustable so that by giving the upper surface a small tilt, localized interference fringes can been seen between them. If the upper surface, the one under test, is flat, the fringes will be parallel straight lines, and any deviation measures the departure from flatness of the surface. The fringes can be photographed and the defects measured to an accuracy of $\sim$200 Å. The division of amplitude here is at the tested surface, usually not coated, so that the fringe contrast is low, but generally sufficient for the test.

A variation on this is to coat the surface under test with a multilayer reflecting film, in which case *Tolansky fringes* are seen, analogous to the very sharp *Fabry–Pérot fringes* to be described below. The fringes then show a contour

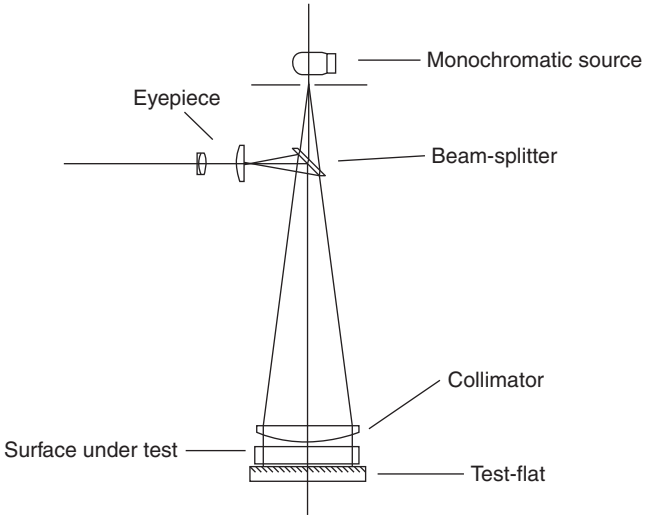

Figure 7.3 The Fizeau interferometer. This is a standard laboratory and optical workshop instrument for testing the quality of an optical surface. Traditionally a low pressure mercury lamp with a didymium glass filter (see Chapter 12) provides monochromatic green light at a wavelength of 546.1 nm. The size of the test-flat varies, but 10 cm diameter is not unusual.

map of the gap thickness, and variations of a few angstroms can be measured. In freshly cleaved crystal surfaces such as mica, 'steps' in the contours due to the edge of a single molecular layer are visible and can be photographed.

7.8 The Michelson interferometer

Albert A. Michelson (1852–1931) was a Lieutenant-Commander in the US Navy who, as an instructor at the US Naval Academy, Annapolis, came to the notice of the US Naval Observatory for his refinement of Foucault's method for measuring the speed of light. He is best remembered, however, for his invention of the interferometer, which is in universal use today. Its applications are manifold:

(1) as a means of measuring wavelengths and comparing them with a mechanical standard of length such as a platinum metre-bar,
(2) as a *Fourier multiplex spectrometer*[6] with a very high resolution–luminosity product,
(3) as a famous test of the existence or non-existence of the luminiferous aether,
(4) as a possible detector of gravitational radiation,[7] and
(5) with fibre-optic arms, as a general remote sensor and measurer of small displacements and disturbances.

In its basic form a collimated beam of monochromatic light from an extended source is directed at a beam-splitter, a pair of glass plates accurately the same thickness and generally made by cutting a single plate into two after polishing it to constant thickness to an accuracy of $\pm\lambda/10$. A reflecting coating is applied to one inner face before the two plates are connected.[8] This is to ensure that the paths of the two beams – transmitted and reflected – are the same for all wavelengths. After reflection and recombination at the beam-splitter the transmitted light is focused on to a plane where, at normal adjustment, an image of the source is seen, with a series of concentric rings superimposed, with angular radii corresponding to path-differences which are an integer number of wavelengths.

The path-difference can be changed by moving one of the two reflectors, and as it increases the ring system will contract, with more rings appearing

[6] Described in Appendix C.
[7] Not to be confused with *gravity waves* which are a meteorological phenomenon caused by a disturbance of the atmosphere.
[8] In early versions the two plates were separate, making an adjustment of parallelism another chore for the operator.

at the edge. If a small hole is made at the centre of the ring system, light will be transmitted, and if the path-difference changes uniformly, the intensity of the transmitted light will vary sinusoidally. One cycle of the variation will correspond to one wavelength change of path-difference.

Therein lies the chief problem of making an interferometer. The path-difference can by changed in several ways. As it is changed, the alignment of the moving mirror must not deviate by more than about 0.2 seconds of arc, or a tilt of 50 nm at one edge. To do this while moving the mirror along a 20 mm change of path was the challenge faced by early researchers and early on was thought to be a mechanical impossibility. Typical solutions to the problem are shown in Figure 7.4. The cube-corner is a set of three reflectors all mutually

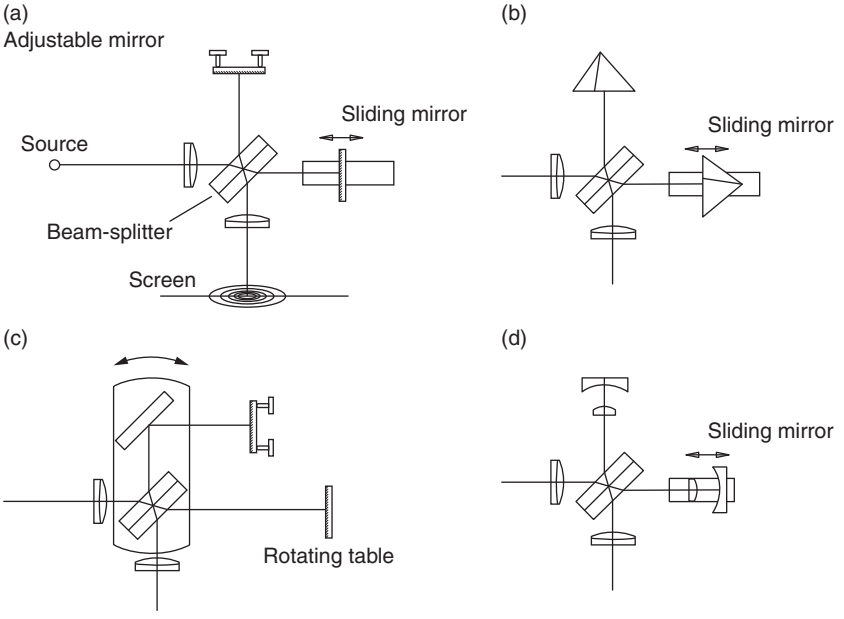

Figure 7.4 The Michelson interferometer, in its various incarnations. The original device is shown in (a). For interference spectroscopy the basic practical problem, that of changing the path-difference without losing the optical alignment, was solved in many ways, chiefly those shown. The use of cube-corner reflectors (b) tends to reduce fringe visibility by changing the polarization characteristics differently in the two arms. The 'tilting Michelson' (c) gives satisfactory fringes but at high resolution requires over-sized plane mirrors in the tilting arm. The cat's-eye reflectors (d), being concentric spherical mirrors, suffer from unequal spherical aberration, again reducing the fringe visibility. All four variants have been used successfully in commercially made instruments.

perpendicular, and it is easy to show that any incident beam is returned in exactly the opposite direction.

The cube-corners may be mirrors, but are usually of glass or quartz, carefully worked to make the angles precisely 90° and relying on total internal reflection.

The so-called 'cat's eye' reflectors are concentric spherical mirrors reflecting any incident ray back the way it came. The 'tilting' interferometer relies on the rigidity of the rotating frame with its two accurately parallel mirrors, although no particular accuracy is required of the hinge.

For a given rate of change of path-difference, the frequency of the intensity fluctuation depends on the wavelength of the light, and this is why the device works as a spectrometer: it frequency analyses the output signal. Its light-thoughput (étendue) is hundreds of times greater than that of a grating spectrograph of the same size, and this makes it very convenient for on-line analysis in the chemical industry. The theory of its working is given in Appendix C.

7.9 The Fabry–Pérot étalon

This is the most widely used of all optical interferometers, being very versatile in its applications and the most manageable field instrument. It is a *multiple-beam* interferometer, as opposed to previous interferometers which were strictly two-beam devices It was the invention of Charles Fabry (1867–1945) and Alfred Pérot (1863–1925) of the University of Marseille. Traditionally the étalon comprised two flat quartz plates separated by three spacers of equal thickness disposed around the edge of the gap. The inner surfaces of the quartz plates are polished to a high degree of flatness, to a tolerance typically of $\pm\lambda/100$ of green light. What is important, however, is not that the individual surfaces be flat but that the *gap* be constant when aligned. Older, hand-polished étalons may therefore have to be put together in the correct orientation and the two plates are generally marked with a pair of arrows pencilled in, with points facing each other when the position is correct.

In the research environment étalons are generally *wrung* together with a fixed quartz spacer between them and gap adjustment is neither possible nor necessary. The gap-width is chosen beforehand to suit the investigation that is under way.

If monochromatic light is incident in plane waves at an angle θ to the gap surface, the path-difference between wavefronts successively reflected twice before onward transmission is given by:

$$\delta = 2md\cos\theta \tag{7.3}$$

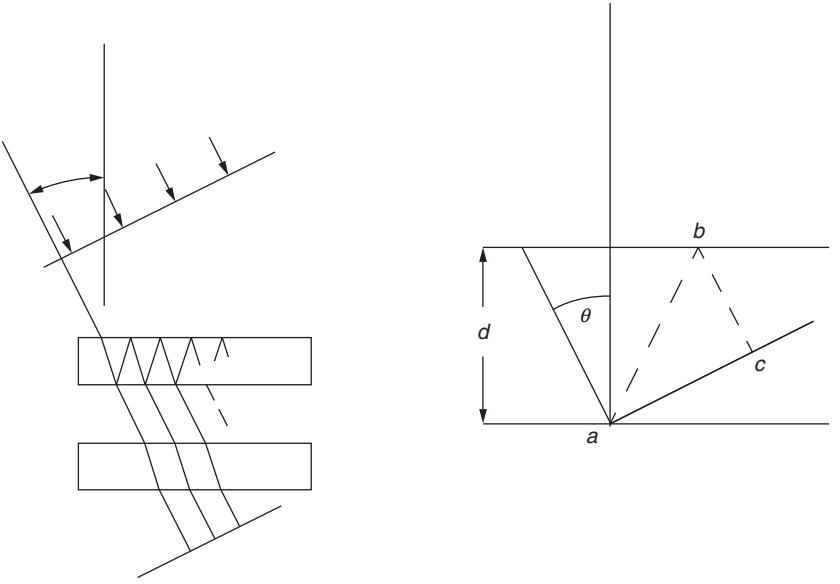

Figure 7.5 The optical path through a Fabry–Pérot étalon. Again, elementary Euclidean geometry will show that the path difference between two rays, one transmitted, the other reflected twice along the path $a \to b \to c$, is $2nd\cos\theta$, where n is the refractive index in the gap. If this path-difference is an integer number of wavelengths, amplitudes will add to form the wavefront $a \to c$ and amplitudes will continue to add for every successive double-reflection. The net result is that the transmitted power is very great; but at other angles, where the wave crests do not reinforce in exactly this way, the amplitudes rapidly change phase and the net effect is that they cancel out. Consequently, at particular angles, given by equation (7.2), there will be sharp maxima of transmitted light with near-darkness between. The result, when the light is brought to a focus, is a concentric series of rings, the so-called Fabry–Pérot fringes. Each ring represents a different order of interference, the order deceasing as the ring diameters increase.

where d is the gap thickness and m is an integer, the order of interference. There are multiple double-reflections, of exponentially decreasing amplitude and constantly increasing phase-difference, and these amplitudes all add vectorially to contribute to the final transmitted amplitude. If, and only if, the path-difference between adjacent beams is an integer number of wavelengths, the amplitudes are all aligned and the resultant is a maximum. Otherwise the vectors, all having different phase-angles, roll up into a tight scroll with a very small resultant. The end of all this is that for certain angles θ, there is almost perfect transmission of the light; otherwise it is all reflected.

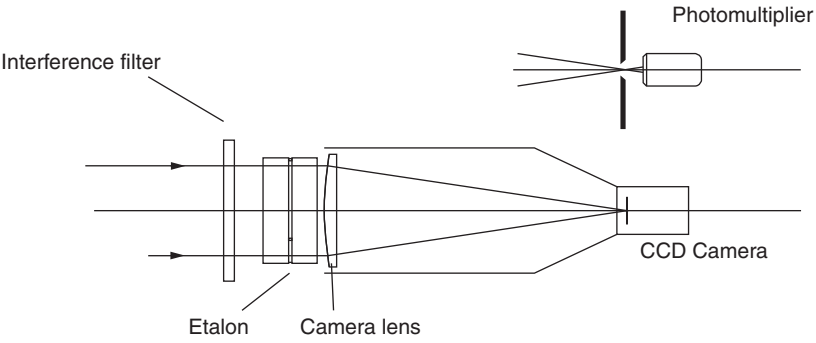

Figure 7.6 The Fabry–Pérot spectrograph. This has an efficiency (RL product) far in excess of any grating spectrometer, and since throughput (L) can be traded for resolving power (R), it serves several purposes. (1) It can be used for very high resolution spectrometry, showing the hyperfine-structure in atomic spectra. (2) It can be used at moderate resolution to measure very feeble sources of light, such as spectral lines in the night airglow and aurorae. (3) It can measure very short duration spectra, such as come from sparks or explosions. Its limitation is its very small free spectral range, $\Delta\lambda = \lambda^2/2d$, where d is the gap thickness. An interference filter, complete with dye filters as 'side-band cutters' must be used to select a pass-band of a few angstroms for the étalon itself. The spectrum, with the light dispersed radially outwards, may be recorded on a CCD chip at the focus of the camera lens. Alternatively the light may be directed towards a small circular aperture on the optic axis beyond which is a photodetector. In the latter case some factor must change in order to present different wavelengths in succession at the aperture. This may be the change in refractive index with increasing gas pressure in the gap, or the gap itself may be changed by piezo-electric transducers with a sawtooth ramp waveform applied.

There is an interesting but neglected comparison of the action of the étalon and a diffraction grating. A point source of monochromatic light shone through an étalon has an infinite number of reflected images of steadily decreasing amplitude along a perpendicular to the plane of the reflectors. These images are all coherent and the resultant amplitude transmitted bears comparison with a beam from a diffraction grating with slits spaced at twice the étalon gap, with the diffracted light being at a very high order. The formal mathematics is the same, with only the exponential decrease in amplitude of the light from more distant slits making any difference.

Thus, when light from an extended monochromatic source (a screen illuminated by a laboratory spectral lamp for instance) passes through the étalon, a lens will focus on to a screen a series of very sharp concentric rings of light, called Fabry–Pérot fringes, with effectively no light transmitted in any other directions. In the interferometer equation above, d, n and λ are constant and

θ is a function of m, the order, which incidentally *decreases* as the angle θ increases. The diameters of the rings depend on the focal length f of the lens and are well described [9] by the equation:

$$D^2 = 8f^2\left(1 - \frac{m\lambda}{2nd}\right) \tag{7.4}$$

Alternatively the intensity transmitted can be described as a function of a useful parameter δ by the *Airy formula*:

$$I(\delta) = \frac{I(0)}{1 + \frac{4r}{(1-r)^2}\sin^2\left(\frac{\delta}{2}\right)} \tag{7.5}$$

where δ is given by

$$\delta = (2\pi/\lambda)2d\cos\theta$$

Again, one may select any of the variables, λ, d or θ in the parameter δ to plot against intensity, keeping the others constant.

As a spectrograph the étalon is useful only over a small *spectral range* because a wavelength λ passing through in order m will show a ring of the same diameter as wavelength $\lambda + \Delta\lambda$ at order $m - 1$. The free spectral range (FSR) $\Delta\lambda$ is given by:

$$\Delta\lambda = \lambda/m = \lambda^2/2d \tag{7.6}$$

and to avoid ambiguity in the measurements, it will probably be necessary to filter the light entering the instrument to restrict the transmitted spectral bandwith to less than the FSR. Such a filter is called an order-sorter. An order-sorter may be an interference filter, it may be another étalon with a gap narrower by a factor of 30 (and hence 30 times the spectral range), or it may be a spectrometer with a wide exit slit allowing the required spectral range to pass. Which one is chosen depends on the nature of the spectrum and on the resolving power required.

A figure-of-merit for a Fabry–Pérot spectrograph is its *finesse*. This is the ratio of the FSR, $\Delta\lambda$, to the half-width of the Airy function, which is the shape of a monochromatic spectrum line when it passes. The ratio of the FSR to this half-width $\delta\lambda$ is a property of the instrument itself and is called the *finesse*. It is the number of spectral elements it is possible to resolve in the free spectral range. For all practical purposes the finesse is given by

$$F = \frac{3}{1 - r} \tag{7.7}$$

[9] Provided the approximation $\cos\theta = 1 - \theta^2/2$ is adequate.

where r is the reflection coefficient of the coating on the plates. In practice, with multilayer coatings, r is typically $\sim$0.97 at the design wavelength, and should give a finesse of $\sim$100, but this assumes that the gap is perfectly flat. In large étalons this is not so and a working finesse in the region of 40–50 is usual. There are fishermens' tales about the finesse of a particular étalon, but really high finesse only comes with a small aperture.

Multilayer coatings are efficient only for a region of a few hundred angstroms around the design wavelength, so that new coatings may be needed for a new experiment. In a teaching laboratory environment, aluminium or silver coatings may be used, renewed as necessary and useable over the whole wavelength range. The finesse of such coatings is unlikely to be greater than about $\sim$15−20.

7.9.1 Uses of the Fabry–Pérot étalon

The device was intended initally for very high resolution spectroscopy and with a finesse F and and a gap d the resolution, $\delta\lambda$, defined as the full width at half maximum (FWHM) of a monochromatic line is

$$\delta\lambda = \frac{\lambda^2}{2Fd} \tag{7.8}$$

and for example, with a finesse of 30 and a gap of 6 mm the resolution at 4200 Å (blue light) is 0.0001 Å, enough to resolve the isotopic hyperfine-structure of uranium and so to measure the isotopic ratio. Gaps of anything up to 1 metre have been attained and when the device is the core of a gas laser this is not unusual. However, when the full aperture of perhaps 4 or 5 centimetres is to be used, alignment with the large gap-widths needed for very high resolution spectroscopy needs high precision adjustments, and these are the province of the dedicated expert.

7.9.2 The demountable étalon

This is the traditional form of the étalon. The two plates, usually of quartz, are reflection coated and are then put together with their three spacers held in a metal mounting ring, the whole assembly held together by three springs pressing on the upper plate over the three spacers.

Traditionally, watchmakers' 'jewels', i.e. rubies, were used as spacers, each carefully ground and polished to have the same thickness. More frequently three ball-bearings, held loosely in a retaining ring at 120° intervals around the edges make a spacer. Ball bearings, by their mode of manufacture, are all of

equal diameter [10] and because of this property they are ideal as interferometer spacers. It was usual to take a box of ball-bearings and by trial-and-error to select three of the same diameter to within a fraction of a wavelength. This relieved the need for undue compression of the quartz plates themselves when the final adjustment of the gap parallelism was being made.

Adjustment of a demountable étalon

In the laboratory the adjustment is done visually. The procedure consists of initally adjusting the gap for parallelism by observing the multiple images of a distant object through the étalon. When these fuse together, the étalon is mounted in front of an extended monochromatic light source, and then by looking not *at* it, but *through* it with the eye relaxed to infinity – which is where the fringes will appear to be – a concentric series of circles will be seen.

Remember that the eye is observing through only a few millimetres of the étalon area and the fringes will be sharp although the gap is not yet uniformly flat. By moving the eye about the area of the plates, the rings will expand or contract, and the final step in adjustment is to change the pressure over the three spacers until the rings are of the same diameter everywhere.

If the étalon has been properly designed the adjustment will hold in an open laboratory (or an eclipse site!) for 15–30 minutes but should be checked as often as possible during an experiment.

7.9.3 The étalon in metrology

A change in gap width of $\lambda/2$ will shift the ring system through one order. A change in gap one-thirtieth of this will move a monochromatic line through one resolved element. A change in gap of as little as 0.01 Å (10^{-12} m) is detectable by observing the change in intensity at the steepest part of the line profile, and the gap width can be controlled to this accuracy, provided of course that suitable precautions are taken.

7.9.4 Edser–Butler fringes

When white light is shone through an étalon only those wavelengths which are an exact fraction of the gap width are transmitted. The transmitted light, analysed by a spectrograph, will show sharp lines, equally spaced in *wavenumber* (which at a large étalon gap are practically the same as equally spaced wavelengths). These can be used as calibration marks for a prism spectrograph for

[10] The steel balls used in ball-bearings are not necessarily spheres: all that is required is that they be surfaces of constant diameter.

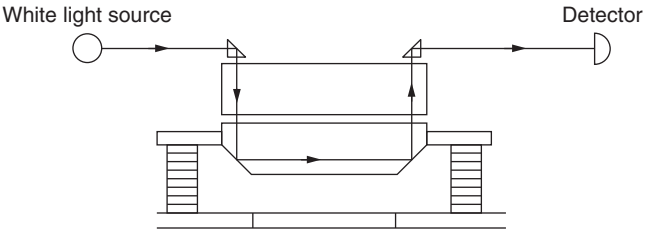

Figure 7.7 Ramsay's optical servo. This is a way of measuring and controlling the gap parallelism in an étalon. It shows the chief-ray of a beam of white light passing twice through the perimeter of the étalon gap. The first passage reduces the light spectrum to Edser–Butler fringes, as only wavelengths which are a half-integer fraction of the gap width can pass. The second transit of the gap requires the second gap width to be exactly the same as at the first passage, with a sharp fall in intensity if there is any misalignment. A small vibration of the gap width (a few angstrom units) will give a sinusoidal signal at the detector and its phase reverses as the gap moves through the correct alignment. This provides a delicate measure of the equality of gap width and can be used to correct the error. In practice, two such beams are used, orthogonal to each other, to ensure perfect gap control in all directions.

example, where the dispersion is non-linear. There is a particularly ingenious application of Edser–Butler fringes by J. V. Ramsay, relying on the fact that full transmission through two successive étalons can only happen if the two gaps are precisely equal. Ramsay's method was to shine white light through the edge of an étalon, reflect it twice and then shine it through the opposite edge. Only if the gap were exactly flat would there be complete transmission. A vibration of one étalon plate through an amplitude of a few angstroms, induced by a piezo-electric ceramic, changes the transmission enough for it to be measured and a correcting signal applied to the same ceramic to hold the gap flat with a precision of $\lambda/1000$.

7.10 The Mach–Zehnder interferometer

This is an alternative to the Michelson interferometer.

Like the Michelson and the Fabry–Pérot interferometers, it works by division of amplitude. Following separation at the beam-splitter, the two optical paths are separate and on recombining normally show zero path-difference when properly adjusted. If some disturbance is put in one path the result is a series of fringes and, for example, a glass plate will reveal point-to-point variations in its optical homogeneity or thickness. As another example, if one

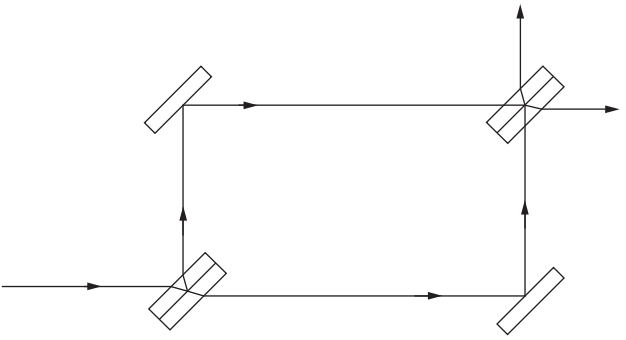

Figure 7.8 The Mach–Zehnder interferometer. This is called a 'four-port' interferometer, as two independent sources can be said to feed light through it, and two detectors can receive light at the output end. This interferometer, as well as being an alternative to the Schlieren system of photography, raises several interesting possibilities for comparing the coherence of different sources. The addition of one or more 'trombone sections' in either path makes the interferometer a Fourier spectrometer. It has many other applications, especially when the two arms are single-mode optical fibres.

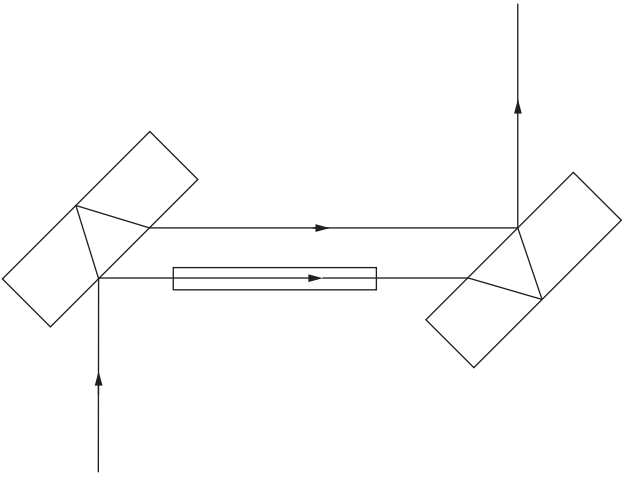

Figure 7.9 The Jamin interferometer. This apparently simple device is a satisfactory alternative to the Rayleigh refractometer, not least because it produces fringes which are easily accessible and fit for photography. It requires a very thick slab of glass of high homogeneity, and relies on geometry to separate the two beams. It is analogous to the Mach–Zehnder interferometer in the way it separates then recombines the beams. Like the Rayleigh refractometer, its principal application is in the measurement of refractive index and dispersion in gases and vapours by the fine measurement of fringe shifts.

arm passes through a wind-tunnel, the changes in refractive index caused by variations of air pressure around a model can be photographed. In this mode it is analogous to a Schlieren system, with the advantage of allowing quantitative measurement of the pressure changes.

The path-difference can also be changed by introducing 'trombone extensions' into one or both paths and in this mode it is a Fourier spectrometer used, like the Michelson, chiefly for infra-red spectrometry.

There is a variant on the idea, invented independently by Jules Jamin (1818–1866) a professor at the École Polytechnique, in which the two beams are separated at the surface of a very thick block of glass and emerge parallel but separated enough that a phase-changing device can be placed in one beam. In this sense it is similar to the Rayleigh refractometer except that it works by division of amplitude rather than division of wavefront.

7.11 The Sagnac interferometer

This is superficially like the Mach–Zehnder interferometer but with the second beam-splitter replaced by a mirror. The two beams then traverse the same optical path in opposite directions and recombine at the original beam-splitter, nominally at zero path-difference. However, if the device is rotating in its own plane there will be a small time-delay in one beam and a corresponding advance in the other, and the result is a small but measurable phase-shift between the two recombining beams: in other words, the device detects absolute rotation. It can be used as a sort of hypersensitive gyrosope because any change of direction causes a small temporary phase-shift as it changes. Like the Michelson and the Mach–Zehnder interferometers, the Sagnac can be constructed using a single-mode optical fibre as the optical path. It can be shown that the rotation-induced phase-shift in the Sagnac interferometer is proportional to the area enclosed by the optical path and a long coil of optical fibre, with its low attenuation, can encompass a large area in a small space. A path-length of several kilometres is typical, encompassing, when wound on a spool, an area of about 100 square metres.

7.11.1 Single mode optical fibres

Briefly, these are waveguides for optical frequencies. A thin (a few micrometers diameter) glass of high refractive index is surrounded ('clad') by another glass of lower refractive index, so that light is differentially refracted or totally internally reflected and propagates down the fibres as a single wavefront.

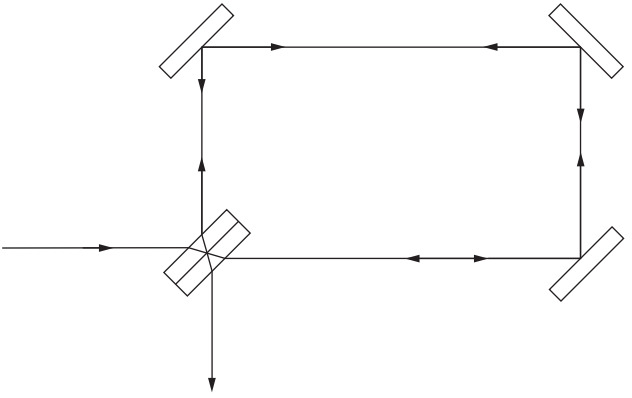

Figure 7.10 The Sagnac interferometer. This resembles a Mach–Zehnder interferometer, but one beam-splitter has been replaced by a mirror, so that the split beams enter and pass round the loop in both directions before recombining. The device becomes interesting when it is rotating in its own plane because then the light takes measurably longer to complete the loop in the direction of rotation than in the contrary loop, and this implies the detection of absolute rotation. In conjunction with a laser input the device is a standard part of an inertial navigation system. With a laser in one arm, it becomes an optical resonator.

Glasses with very low absorption (less than 3 dB/kilometre) in the infra-red are used and these allow very long optic fibre cables to be used for information transmission. In interferometry, the use of fibre cable as the optical arms allows remote sensing of various difficult things such as temperature in a reactor core. Both arms may be several metres or even kilometres long, so long as they both have the same optical length to preserve near-zero path-difference. Anything which affects the optical properties of the fibre can be measured, such as small changes in refractive index induced by temperature changes, and this will show as a phase change, i.e. a fringe-shift at the beam-splitter.

7.12 Fine displacement measurement

Interferometers play their part, but are by no means the only methods of making measurements at sub-angstrom levels. The étalon already mentioned, with one plate fixed and the other suspended by a pendulum 20 cm long, will move by 1 nm if it is tilted by 2×10^{-10} radians. Such a tilt is less than that in daily Earth-tides, the local distortion of the Earth's surface by the passage of the Moon, or to the stress-tilting of the Earth's surface by an impending earthquake, or to the distortion produced by volcanic magma filling a

subterranean chamber before an eruption.[11] However, such fine measurement can be achieved with the same sensitivity by an optical lever such as was used in an old-time galvanometer. An image projected on to a screen after reflection can be stabilized to a small fraction of a millimetre by a pair of photo voltaic or photoconducting sensors separated by a few micrometres. Any inequality of the two signals caused by a tilt or other displacement is amplified and used as a correcting signal – a voltage applied to a piezo-electric transducer for example – to restore the device to its original position. The strength of the correcting signal then measures the original displacement.

However, this is not the whole story. Competing with both these optical methods are devices using capacitative coupling. A tiltmeter can be made from a plate suspended as a pendulum between two fixed plates. The arrangement comprises a pair of capacitors and an a/c bridge will measure their difference in capacity – and hence the changes in the two gaps – as the tilt changes the position of the suspended plate. The net accuracy is about the same as with the other two methods.

7.13 Optical resonators

These are more sophisticated applications of interference, and involve the establishment of optical standing waves which may be maintained, as in a laser, by optical pumping. The Fabry–Pérot étalon is one such. The now-traditional HeNe laser is the earliest example of a sustained resonance, but such resonators now include optical fibres, sometimes toroidal, and small glass spheres with so-called 'whispering gallery' modes of oscillation, sustained by optical coupling (by frustrated total internal reflection) with external laser beams. At the time of writing, optical whispering galleries are still in the research phase and have yet to find practical applications.

[11] As an anecdote, the author once measured the twice-daily tilt of the Schuster laboratory in Manchester University caused by the weight of rush-hour traffic on adjacent Upper Brook Street.

8

Electro-optical effects and their practical uses

8.1 Introduction: polarized light

James Clerk Maxwell (1831–1879) discovered the first unified field theory when in 1865 he showed, drawing on Faraday's experiments, the Biot–Savart law and other results, that interacting electric and magnetic fields and forces could be described by the same set of equations. Oliver Heaviside (1850–1925) reduced the twenty quaternion equations of Maxwell's theory to the four vector equations that ordinary mortals could understand. [1] Maxwell showed that a wave equation could be deduced and that the consequent electromagnetic field propagates in a direction perpendicular to its electric and magnetic field vectors, and at a velocity which could be calculated from two quantities (permittivity and susceptibility) measurable in separate laboratory experiments. The resultant computed velocity was very close indeed to the experimentally measured speed of light. It was an easy conjecture then, that light was electromagnetic radiation. It was a further logical extension to predict radio waves, later confirmed by David Hughes' and Heinrich Hertz's spark-gap experiments.

The mutually perpendicular electric and magnetic field vectors oscillate in amplitude at about 5×10^{14} Hz for red light. If the direction of propagation is the positive z-direction, then the electric and magnetic vectors lie in the $x-y$ plane. The light is said to be *plane polarized* if the angle the electric vector makes with the x-axis is constant in time. We are at liberty to choose the x-axis in any convenient orientation and the electric vector can then be decomposed into its x- and y-components, which are of course coherent.

However, the two components are not necessarily in phase, and if they are not, the light is *elliptically polarized* and the end of the electric vector traces out

[1] Provided, of course, that they had a thorough grounding in vector calculus and differential geometry.

an ellipse as it rotates. If the axes are chosen so that the x- and y-components have the same magnitude and the phase-difference is $\pi/2\,(90°)$, the light is *circularly polarized* and in turn it is *left circularly* or *right circularly* polarized according to which way the vector is rotating. [2]

There is no internationally recognized convention to define the two modes. It depends on whether the coordinate system used to describe the light is left-handed or right-handed. Textbooks and treatises consulted by the author afforded no enlightenment as their authors tended to duck the issue or neglected to define the handedness of their coordinate system. Hereinafter, therefore, we *define* right-hand circularly polarized light as that in which the electric vector at a fixed point, as seen by an observer looking towards the source, appears to be rotating clockwise. [3] If the vector train is frozen for a moment, it appears to be an ordinary cork-screw or right-handed helix. On the whole, optical authors may agree with this, some (but not all) telecommunications engineers [4] may not.

A beam of plane polarized light can be regarded – and it is sometimes convenient to do so – as the superposition of two coherent, oppositely rotating circularly polarized beams.

8.2 Polarization effects

It is important at the outset to distinguish between two effects exhibited by plane polarized light when it interacts with matter.

8.2.1 Birefringence

When light propagates along one of the crystal axes in a non-isotropic crystal, the refractive index depends on the direction of the electric vector with respect to the other two axes. If the plane of polarization is not along one of these axes the two components along those axes will propagate at different speeds and there will be a gradually increasing phase-difference as the light propagates. The light will then emerge elliptically or circularly polarized. Similarly, when

[2] If you are mathematically inclined you can regard the electric vector as a two-component spinor, to be manipulated using Jones calculus.

[3] Which exposes a fundamental and subtle problem: how do you define ‘clockwise’ to an Andromedan?

[4] The satellite receiving dish at Goonhilly Downs, Cornwall, designed to receive TV signals from the Telstar satellite, was formally inaugurated in 1962. The story goes that on commissioning day the switch was ceremoniously pulled – and nothing happened – the circularly polarized radiation from the satellite was of the opposite hand to that receivable by the dish.

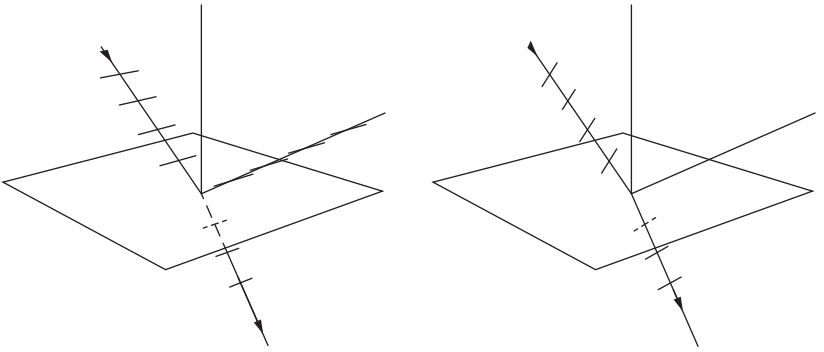

Figure 8.1 Brewster angle reflection and refraction. The intensity of reflected light incident on a surface at the Brewster angle depends heavily on the direction of polarization of the incident light. If the refracted and reflected ray directions are perpendicular and the electric vector of the refracted ray points along the direction of the reflected ray, the amplitude of the reflected ray will be zero. This is what happens at the Brewster angle of incidence.

light is reflected from the surface of any homogeneous transparent material such as glass or water, the reflection coefficient depends (a) on the angle of incidence and (b) on the plane of polarization. If the electric vector is in the plane defined by the surface normal and the incident ray, the reflection coefficient is given by

$$R_1 = \frac{\tan^2(i-r)}{\tan^2(i+r)} \tag{8.1}$$

and if the vector is parallel to the reflecting surface,

$$R_2 = \frac{\sin^2(i-r)}{\sin^2(i+r)} \tag{8.2}$$

where i and r as usual are the angles of incidence and refraction. If there is no absorption, then power that is not reflected is refracted. Notice in particular that $R_1 = 0$ when $i + r = 90°$, so that there is no reflection of light polarized in the plane of incidence when $\tan i = n$, the refractive index. Any light reflected at this angle is therefore plane polarized with its called electric vector parallel to the surface. This angle of incidence is called the *Brewster angle* after Sir David Brewster (1781–1868). For water, the Brewster angle is 53°7′. Brewster angle reflection is a possible method of generating a beam of plane polarized light, but a very inefficient one, since the intensity reflected is less than 15% of the incident intensity. On the other hand, it follows that light polarized parallel to the surface is refracted without loss and *at both surfaces of a glass plate*. This was put to good use in the early days of the HeNe laser, when amplification

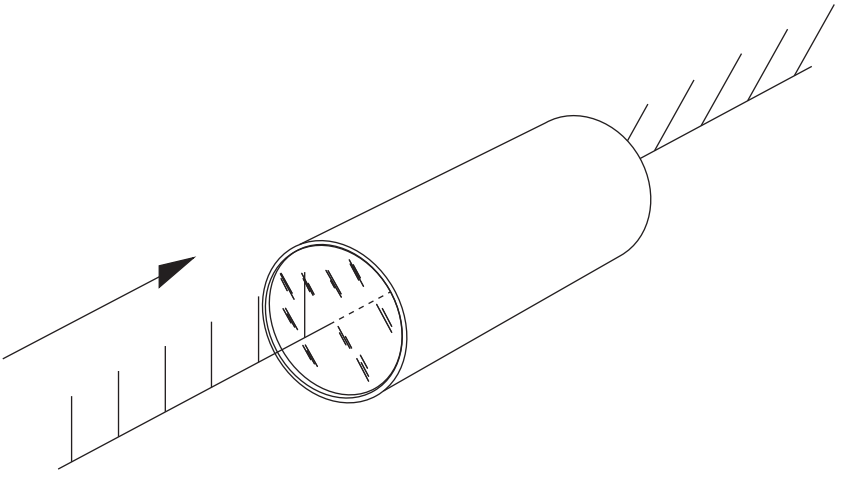

Figure 8.2 Optical activity. This is intrinsic to some materials, and not an effect imposed by applied fields. If a charge at one end of a twisted molecule with an intrinsic electric dipole moment is set vibrating by an electric field, the vibration is transferred to the other end of the molecule where the complementary charge will be set oscillating almost but not exactly parallel to the original. The angle will be small (less than 10^{-2} seconds of arc), but cumulatively the electric vector of the transmitted polarized light will be turned. The direction of twist will depend on which isomer of the molecule is present. The amplitude of rotation does not depend on the polarization orientation of the incident light. Circularly polarized light is unaffected apart from a minuscule change in refractive index which depends on the handedness of the circular polarization.

was low, and the ends of the laser tube were closed with glass discs inclined at the Brewster angle to ensure that there was no loss between the reflectors for at least one plane of polarization.

8.2.2 Optical activity

The propagation of plane polarized light *in any direction* through a chiral medium, generally an aqueous solution of sugar or some other anisotropic molecular species, results in a rotation of the plane of polarization, the rotation angle increasing progressively with depth and depending linearly on the concentration of the solute. The direction of the rotation may be clockwise or anticlockwise (as seen from the source of the light) and for example the sugar *dextrose* will rotate the plane clockwise and *laevose*, also known as *fructose*, will rotate the plane anticlockwise. The two molecules are *chiral*, [5] structurally

[5] From the Greek '$\chi\epsilon\iota\rho$' meaning 'hand', a neologism due to Lord Kelvin.

identical but *isomers*, mirror images of each other. An alternative description of optical activity is to say that an optically active medium has a slightly different refractive index for left-handed and right-handed circularly polarized light.

The rate of rotation also depends on the wavelength and increases with shorter wavelengths. The absorption may also be different for different handedness so that plane polarized light may become increasingly elliptically polarized as it propagates.

8.3 Polarizing elements

8.3.1 Plane polarizers

Traditionally anisotropic crystals were used to make *Nicol prisms* or *Wollaston prisms* from crystals of Iceland spar (also known as calcite), a crystalline form of calcium carbonate, a uniaxial crystal in which the two planes of polarization are refracted at slightly different angles and are separated, either by having one plane of polarization totally internally reflected (Nicol) or by having the two planes refracted at different angles within the prism (Wollaston). These are largely consigned to history except in specialized instruments, and polarized light is generally produced using familiar laboratory materials such as polaroids, which are transparent to one plane of polarization and opaque to the other. In such materials, no matter what light is incident, that which is transmitted is plane polarized.

8.3.2 Wave plates

The other useful and common optical elements are the *quarter-wave plate* and the *half-wave plate*. The quarter-wave plate is used to convert plane polarized light to circularly polarized light or *vice versa* and the half-wave plate turns the plane of polarization through 90°. The most common wave plates are thin sheets of mica, a uniaxial crystal in which the birefringence direction is perpendicular to the cleavage plane. Thus light normally incident is transmitted with a refractive index which depends on the plane of polarization. For a quarter-wave plate, the thickness of the mica is carefully chosen so that light with one plane of polarization is retarded by a quarter of a wavelength with respect to the other. Similarly, a half-wave plate has twice that thickness so that the retardation is half a wavelength. **To achieve the purpose of converting plane polarized light to circularly polarized light, a quarter-wave plate must be mounted perpendicular to the optic axis and oriented so that its slow and fast axes are at 45° to the plane of polarization.** It this is not the case, then

the emerging light is elliptically polarized. The correct orientation is chosen by finding, by visual observation usually, the position when the emerging light is still plane polarized, and then rotating the plate in its own plane through 45°.

There is a small wavelength dependence on the amount of retardation, so that strictly speaking a quarter-wave plate gives *exactly* a quarter of a wavelength retardation for only one wavelength; but because the optical thickness increases for both planes of polarization as the wavelength decreases, the changes are slow and a quarter-wave plate is useful over a region of several hundred angstroms around the exact wavelength. It may be possible to improve its performance if necessary, by tilting it slightly with respect to the optic axis, so that incidence is slightly off-normal.

Such plates are sometimes (but not always) marked with the direction of the 'fast' axis and if plane polarized light is incident with the plane at $\pm 5°$ to the fast axis, a quarter-wave plate will convert it to either left or right circularly polarized light. A half-wave plate will change the plane of polarization by 90°.

8.3.3 Circular polarizers

A common combination is to take a plane polarizer – a sheet of polaroid in practice – and attach it to a quarter-wave plate in such a way that the axes of the quarter-wave plate are at 45° to the polarization axis of the polaroid. The combination is to be found in the polarizing filter made for ordinary CCD cameras and is intended for landscape photography. It relies on the fact that light from the clear blue sky is partially plane polarized, the electric vector being perpendicular to the (great circle) line from the Sun to the scattering place. The polarization is a maximum at a solar elongation of 90° and the effect is to darken the sky, to highlight cumulus clouds and increase the dramatic contrast of the photograph. The combination of polaroid + quarter wave plate is also known sometimes as 'one-way glass' because if a sheet is laid over a metallic reflector, polaroid up, the light reflected from the metal surface is still circularly polarized but since its direction of travel is reversed, the helicity is in the opposite sense. The quarter-wave plate returns it to plane polarized light but now the plane of polarization is perpendicular to the polaroid and the light is not transmitted. If the plate is turned over there is no such effect.

8.4 Laboratory uses of polarized light

There are both qualitative and quantitative uses for a collimated beam of polarized light. The most obvious quantitative use is in *saccharimetry*, the

measurement of the concentration of sugar in solution, and saccharimeters are standard instruments in the laboratories of the brewing and distilling industries. Qualitative uses include the use of the phenomenon of *stress birefringence* or *photoelasticity* in transparent materials to estimate the distribution of stress in engineering components made of materials such as steel and aluminium. Transparent plastic replicas, typically polythene or polycarbonate, are made of the same shape and with the same stresses applied. The test piece is mounted between crossed polaroids and examined in a beam of collimated light to make the measurements. Strictly, quarter-wave plates are mounted after the first polaroid and before the last one so that the specimen is examined in circularly polarized light, and this avoids the effect of different wavelengths showing different amounts of birefringence.

Transparent samples of mineral crystals and some other transparent materials (such as cellophane) which show natural birefringence or optical activity can also be examined between crossed polaroids in this way.

Many anisotropic crystals exhibit the *piezo-electric effect* and expand or contract in the direction of an applied electric field. Barium titanate is pre-eminent among them. However, if the crystal is transparent it will exhibit stress birefringence as a consequence of the elastic deformation. This effect is used in some semiconductor materials to achieve acoustic modulation of a light beam. Gallium phosphide is the material of choice in this case.

8.5 Faraday effect

This is named after its discoverer, Michael Faraday (1791–1867). The effect can be called 'induced optical activity'. Being an electromagnetic effect, it works at all wavelengths including radio wavelengths. The 'Faraday rotation' of satellite signals due to the interaction between free electrons in the ionosphere, the Earth's magnetic field and the outgoing and incoming radiation, rotates the plane of polarization of satellite radio signals[6] and allows aeronomers to estimate the electron density along a column of the atmosphere.

The effect is linear and depends on the magnetic field strength, the length of path through the field (found by integration if necessary) and a constant, *Verdet's constant*, which depends on the particular material involved. The equation describing it is:

$$\theta = vBd$$

[6] This is why ground to satellite communication is with circularly polarized microwave transmitters.

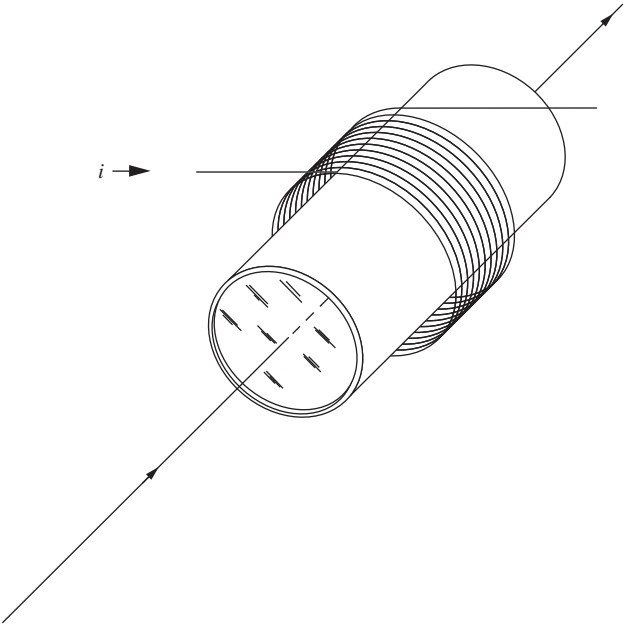

Figure 8.3 The Faraday effect. This is induced optical activity, caused by the effect of a magnetic field on some substances. As with naturally active substances, the effect does not depend on the initial orientation of the electric vector of the incident light.

where θ is the rotation of the plane in radians, v is Verdet's constant (usually in radians per tesla metre), B is the field strength in Tesla, and d is the length of the path in the field.

The effect is heavily dependent on wavelength, increasing at shorter wavelengths, and often Verdet's constant is negative, implying that the plane of polarization rotates anticlockwise as the light propagates through the material. Silica has a Verdet constant $\sim$4 (radians per tesla metre) at 632.8 nm. Flint-glasses have Verdet constants $\sim$7 rad/(T m) and the highest Verdet constant known is in terbium gallium garnet, at -134 rad/(T m). There are transparent materials with near-zero Verdet constants.

8.6 Kerr effect

Light passing through an electric field in an active medium (all media are active to some extent), perpendicular to the electric field vector, experiences a different refractive index if polarized parallel to and perpendicular to the

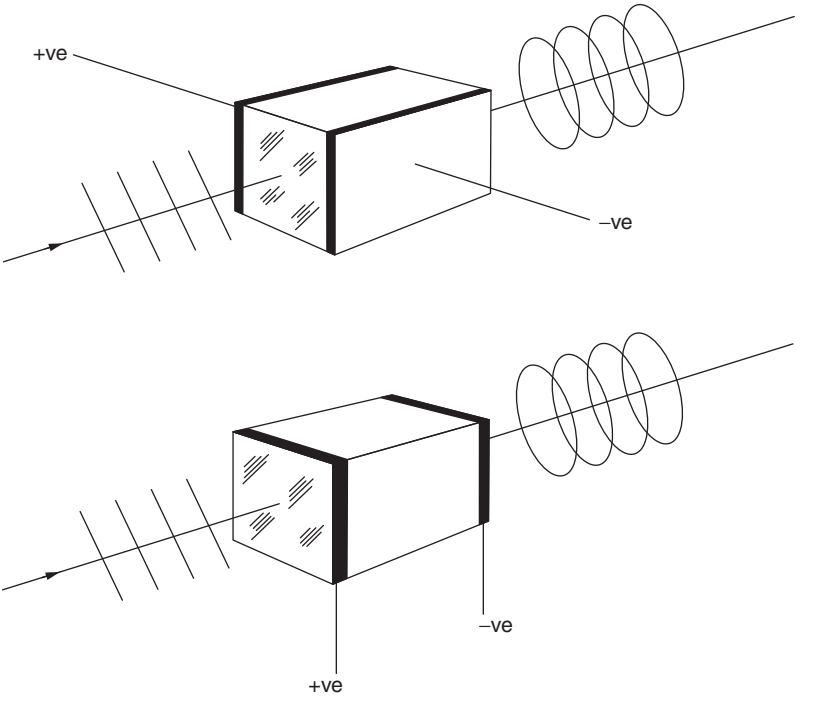

Figure 8.4 The Kerr effect (top) and the Pockels effect (bottom). These are both effects in 'active media' wherein an applied electric field induces a dipole moment in the molecules. The Kerr effect depends on the relative directions of the electric vectors of the applied electric field and the incident plane polarized light. The Pockels effect, on the other hand, is seen only in anisotropic materials, typically potassium dihydrogen phosphate (KDP), and, like the Kerr effect, changes the amount of birefringence. In other words, the crystal becomes a wave-plate.

field direction. The effect is also called 'induced birefringence'. The difference between the refractive indices is described by the equation:

$$\delta n = K.\lambda E^2$$

Where n is the refractive index, λ is the wavelength of light, in metres, and E is the electric field strength, in volts per metre.

The material then acts like a wave-plate and changes the ellipticity of polarized light passing through in the appropriate orientation.

Kerr constants vary widely. Nitrobenzene ($C_6H_5NO_2$) is outstanding and is used in *Kerr shutters*. These comprise cells of nitrobenzene between crossed polaroids, and are 'closed' when no field is present. They are opened

momentarily by a high voltage pulse applied perpendicular to the transmission direction, and the pulse may be as short as a few nanoseconds.

There is another, less well-known *magneto-optic Kerr effect*, sometimes called SMOKE'.[7] Plane polarized light reflected from the polished surface of the pole of a magnet has its plane of polarization rotated. The amount depends on the (complex) permeability, μ, of the material of the magnet pole.

8.7 Pockels effect

This is also called the linear Kerr effect although the basic cause is different. It is found only in anisotropic crystalline materials and potassium dihydrogen phosphate (KDP) is the currently favoured material. When an electric field is applied across two opposite faces of the crystal it becomes a wave plate, and a potential of about 3.5 kilovolts will convert the crystal into a quarter-wave plate. To achieve this, transparent conducting electrodes are necessary (despite the fact that the phrase 'transparent conductor' is almost an oxymoron). Indium tin oxide is the material of choice at the time of writing, but newly discovered materials such as graphene will probably change this. As in the Kerr effect, high shutter speeds are possible, with the limit set by the thickness of the cell and the high dielectric constant ($\sim$6) of the material which reduces the speed of light in the crystal to about 2 inches/ns.

8.8 A practical example of applied polarizing effects

An interesting application of various polarizing effects can be found in a polarizing Michelson interferometer. The light source is a HeNe laser beam, plane polarized with the electric vector parallel to the beam-splitter face of the interferometer. Both reflected and transmitted beams remain plane polarized. One of the two beams is passed through a quarter-wave plate with its axis at 45° to the plane of polarization and the beam is thus converted to right-handed circularly polarized light. On reflection the beam is converted to left-handed circularly polarized light and the same quarter-wave plate changes it back to plane polarized light, but now with the plane of polarization perpendicular to the original direction.

The two beams recombine at the beam-splitter and the combination emerges, still coherent but non-interfering, as their electric vectors are mutually perpendicular. There is also a phase-difference equal to $2\pi/\lambda$ times the

[7] **S**urface **M**agneto-**O**ptic **K**err **E**ffect.

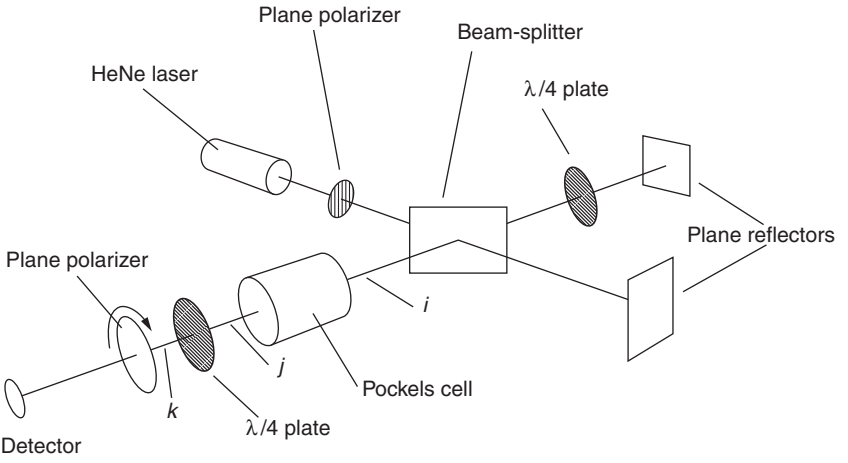

Figure 8.5 The optical screw. A plane polarizer ensures that light from the laser meets the beam-splitter with the electric vector parallel to its surface. A $\lambda/4$-plate in one arm circularly polarizes the plane polarized light and the second passage, following reflection, restores plane polarization, now perpendicular to the original direction. On recombining, the two beams at i have mutually perpendicular polarization planes. An a/c signal on the Pockels cell puts a small cyclic retardation on one of the beams at j. A second $\lambda/4$-plate converts the two beams to right-circular and left-circular polarization and, being coherent, the two add to give plane polarized light at k, but now with the plane of polarization depending on the path-difference and phase-difference in the interferometer. The Pockels cell signal means that the polarization plane vibrates slightly in direction. A final plane polarizer can be turned to extinguish the beam, and the angle of turn depends on the interferometer path-difference. Not only that, but near the extinction point the intensity vibrates in concert with the signal sent to the Pockels cell, so that the phase flips 180° as the polarizer angle passes the extinction point. This gives an error signal which can be used to measure or control the path-difference of the interferometer. One complete turn of the final plane polarizer corresponds to a path-difference change of one wavelength of laser light.

interferometer path-difference, so that the emergent recombined beam is elliptically polarized.

The beam is now transmitted through another quarter-wave plate which, suitably oriented, changes one component to right circularly polarized and the other to left circularly polarized light. As they are coherent the phase-difference has been preserved and the beam finally emerges as plane polarized light, with the plane of polarization depending on the phase-difference, and rotating through 360° when the path-difference changes by one wavelength. The plane of polarization thus measures the path-difference in the interferometer.

If a Pockels cell is put into the emerging beam immediately after the beam-splitter, it can retard one of the two beams relative to the other, so that the angle of the final plane polarized light is changed. If the Pockels cell voltage is alternating, the final polarization direction will vibrate at that frequency and a further plane polarizer followed by a photocell detector will give an alternating current output. The important thing now is that the *phase* of this alternating current will flip by 180° as the polarizer passes through a minimum of transmission: in other words the output is phase sensitive and can be used as an error signal. This signal can be used to hold the final plane polarizer exactly at the extinction direction, thus measuring the path-difference, or it can control one of the interferometer mirrors to hold the path-difference constant [8] despite outside disturbances. Or the final polarizer can be rotated at constant speed and a servo-mechanism can then be made to change the path-difference accordingly. With a stepper-motor to control the polarizer there is an important application in interference spectrometry, a subject dealt with in Appendix C.

[8] Constant to aboout $\lambda/100$ of the laser wavelength.

9

Microscopes and projectors

Microscopes and projectors have this in common: they illustrate the universal requirement that the passage of light through an instrument must be closely controlled. Neither is likely to be improvised by the reader, but one should understand the way they work.

9.1 Projectors

In this section we deal chiefly with the *diapositive* projector, the old-time 'magic lantern' as our Victorian forefathers knew it, which projects images of transparencies on to a large screen for public viewing. Two things are necessary for its design:

(1) it must project a clear sharp image of the transparency on to a distant flat surface, and
(2) it must illuminate the transparency uniformly with light – there must be no dark corners to the picture.

The ingredients for this are a bright source of light, ideally an extended source, but failing that either an array of white LEDs or a filament lamp with a series of coiled filaments as shown in Figure 9.1, arranged so that the space between parallel coils is equal to the width of the coil.

Behind the filament is a spherical mirror placed so that the filament is at its centre of curvature and slightly offset laterally so that the images of each coil fall between the coils themselves. In front of the filament is a condenser lens, not necessarily of high optical quality, often moulded rather than ground and polished, usually with non-spherical surfaces. This should transmit light uniformly over its aperture to form an image of the lamp at the pupil of the projection lens. This will ensure that any light received by the condenser necessarily

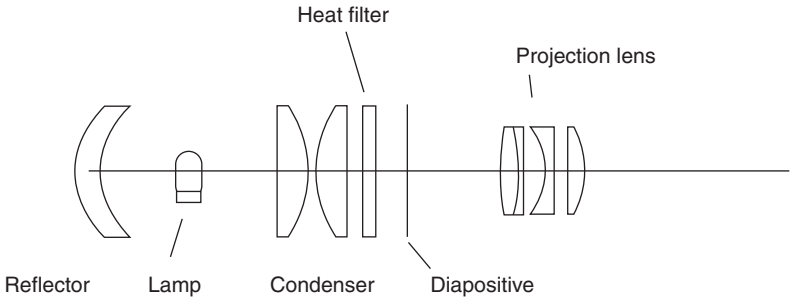

Figure 9.1 The passage of rays through a slide projector. The reflector re-images the lamp on to itself. The condenser images the lamp (and its reflection) on to the pupil of the projection lens, meanwhile filling the slide uniformly with light.

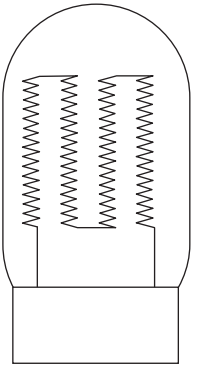

Figure 9.2 The projector lamp. This is the classic design, with spaces between the filaments, to be filled by the image from the reflector. The resulting square of light is intended to give uniform illumination to the diapositive that is to be projected.

finds its way to the projection screen. Following the condenser is a heat filter, a lens which will absorb as much as possible of the infra-red radiation from the lamp and will protect the diapositive which would otherwise absorb the radiation and overheat.

The projection lens is of moderate photographic quality, with the usual attributes of a flat field and freedom from distortion. A high relative aperture is not required as it conflicts with the need for a large depth of focus on the object side. This is to allow for the fact that diapositives are often projected without having been mounted between glass and are consequently not perfectly flat.

There is a more subtle point about slide projection, less appropriate since the advent of zoom lenses, but nevertheless not to be ignored. If a photograph of a landscape is taken with the traditional 50 mm lens in a 35 mm camera and subsequently projected with the commonly used projector lens of 85 mm focal length, the scene will be seen in its correct proportions, that is to say, it will have the same *angular* extent of distant objects as the original scene, only if the distance of the viewer from the screen is 50/85 of the distance of the projector from the screen. In the author's experience, even a photograph of the local high-street may puzzle viewers until they suddenly recognize it when they see it in its proper proportions.

This has its corollary in considering the following question: how far should you be from your television screen? So far as landscapes are concerned there is no obvious answer to this: it depends on the focal length that the cameraman had chosen. But it goes without saying that wide-angle lenses are to be avoided because they exaggerate the foreground at the expense of the distant scenery, so that it is virtually impossible to give a true rendering of a landscape [1] without viewing the TV screen from a much closer distance than is usual for most people.

9.2 The Fresnel lens

These are lenses of moderate resolution, designed primarily to save weight. They take their name from the lighthouse lenses invented by Augustin-Jean Fresnel (1788–1827), which comprised an arrangement of moulded glass prisms set concentrically to act collectively like a lens, without the huge bulk which would have been necessary in an equivalent glass lens. A set of such lenses was mounted in a cylindrical bronze frame, to project the light into several narrow horizontal beams. The whole cylinder rotated slowly, typically once every 10 seconds, to provide the necessary flashes by which the light was identified.

The modern equivalents are thin lenses of high numerical aperture, with diameters greater than 10 cm and without the bulk that an equivalent glass lens would have. They are made of moulded polycarbonate and comprise a large number of concentric annular grooves, each with its surface inclined at the angle which a glass lens surface would have at that radius. They are excellent concentrators of light and form surprisingly good images, comparable with those formed by a glass lens of similar size and numerical aperture.

[1] There is also the problem of 'bulging' while panning in cinematography, as mentioned in Section 4.2.

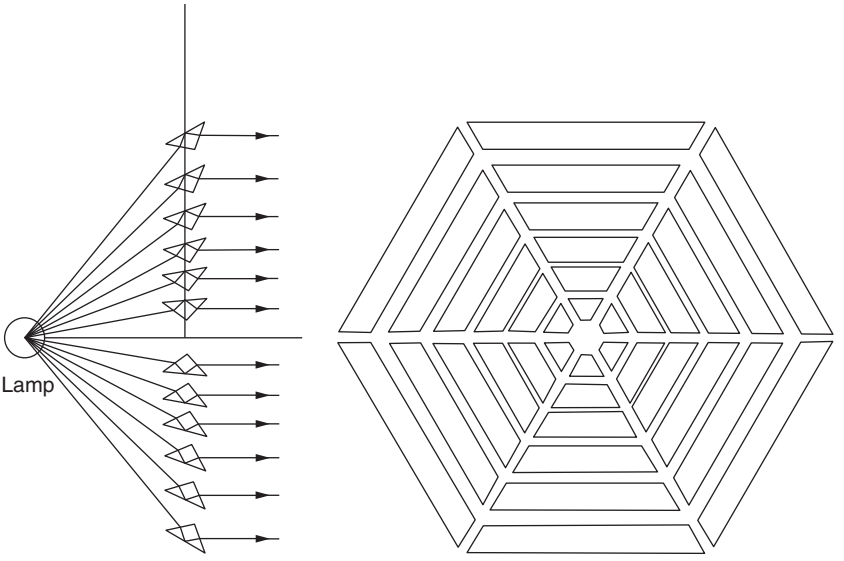

Figure 9.3 The lighthouse lens. This was the invention of Jean-Augustin Fresnel (1788–1827). The construction required no optical polishing work, being constructed from comparatively small cast glass prisms held in a bronze frame. Several of these structures were mounted at intervals around a cylindrical frame, which was then made to rotate slowly to sweep the resulting beams in sequence around the horizon, thus giving the characteristic flashes of the light.

Nevertheless their chief application is as condensers, i.e. as concentrators of light on to a small aperture or projectors of a narrow guidance beam as in the original lighthouse lenses. High optical image quality is not to be expected but the quality is adequate for many purposes.

9.3 Microscopes

9.3.1 The simple microscope

This is a single achromatic lens of very short focal length, in essence a magnifying glass. It is of limited magnification and field, but very compact and portable for field work.

9.3.2 The compound microscope

It is a standard result in diffraction theory (see Chapter 8) that a fine grid will diffract light that passes through it and that the angle of diffraction increases

as the grid becomes finer. For normally incident light the well-known equation describes the effect:

$$a \sin\theta = m\lambda \tag{9.1}$$

where a is the spatial repetition period of the grid, λ the wavelength of the incident light and θ the angle of diffraction. The *order* of diffraction m is an integer $\pm 1, 2, 3...$, positive or negative depending on which side of the grid normal the diffraction takes place.

In order that an optical instrument be able to resolve the grid on an image, it is necessary that *at least* light from orders of diffraction $(-1, 0, +1)$ should pass through the entry pupil of the instrument. If only those orders can pass, then the image of the grid will show a sinusoidal variation in intensity, and this is the limit of resolution of the system. Coarser grids will diffract at smaller angles and the ± 1, ± 2, etc. orders will pass and reveal further details of the grid structure. [2]

Obviously these remarks apply not only to grids, but to any fine-structure on the microscope stage.

This is the core of the theory of the microscope derived by Ernst Abbé (1840–1905). As a consequence, the objective of a microscope is designed to receive and transmit light transmitted by the object under examination, through as wide an angle as possible to collect as many orders of diffraction as possible. There is a further requirement: light must be incident on the object from angles corresponding to these limiting angles prescribed by diffraction. Then zero-order light also arrives at the margin of the aperture and orders $+1$, $+2$,... will be accepted. The microscope therefore will have not only an objective with a large numerical aperture but it will also be illuminated with light from a condenser beneath the stage which emits a similar broad 'pencil' of light.

9.4 Köhler illumination

To obtain best possible resolution it is worth taking some care over the way the object on the stage is illuminated. The basic idea is that beams of *collimated light* should pass through the object stage, with an étendue large enough to fill

[2] The theory is given in detail in Appendix C. The various diffracted orders have amplitudes which are *Fourier components* of the periodic structure of the grid, and this follows from the theorem that the amplitude of the diffracted light, as a function of wavelength and diffraction angle, is the Fourier transform of the amplitude distribution in the object plane.

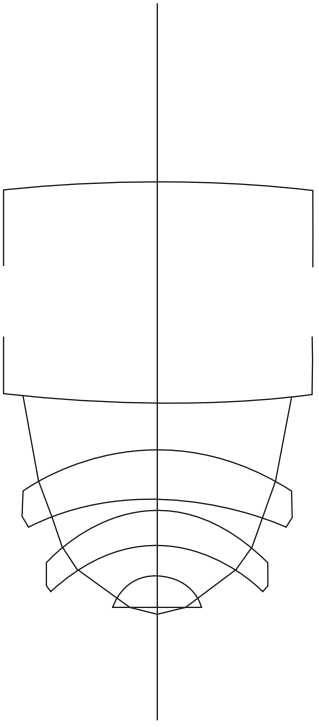

Figure 9.4 The principle of the microscope objective. A near-hemispherical lens accepts light from the specimen at the highest possible numerical aperture. One or more positive aplanats increase the focal ratio (reduce the numerical aperture) and the following lens group deals with the chromatic and other aberrations of the image. The important thing for maximum resolution is that the objective be filled with *collimated* light (and not just scattered light from the specimen) at all possible angles. The condenser below the stage is an essential part of the design.

the whole instrument with light. By étendue here, we mean that the beams are wide enough to illuminate the whole object, and vary in obliquity enough to fill the objective lens. This of course is an instance of the adage that to use an optical instrument properly, you must fill the aperture and field with light.

An extended source of light at $-\infty$ would do this, but that is obviously impractical. Instead, the light source – a frosted filament lamp or an LED – is focused by a condenser lens (or concave mirror) on to the rear focal plane of the collimator below the stage, with a magnification large enough that the image fills the field of the collimator. The collimator then sends collimated light through the object in all directions and simultaneously images the condenser aperture on to the object. This arrangement is similar to that of a projector and ensures:

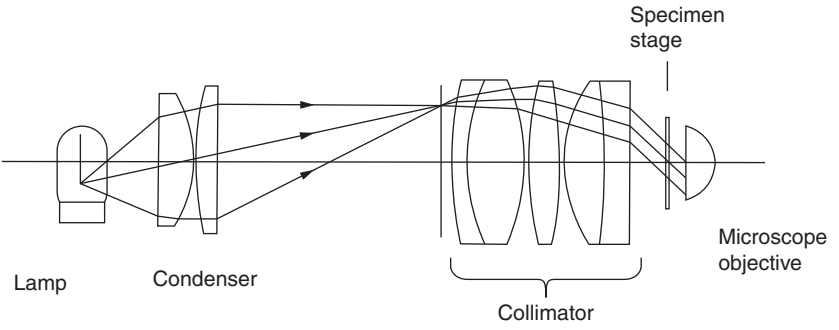

Figure 9.5 Köhler illumination. The carefully designed condenser–collimator system below the specimen stage sends collimated beams of light through the specimen so that the objective may capture as many diffracted orders as possible from the fine detail in the specimen.

(1) that the object is evenly illuminated, and
(2) that coherent wavefronts from every point on the object reach all parts of the objective lens, thus giving the maximum possible spatial resolution on the eye (or camera).

This is called *Köhler illumination* after August Köhler (1866–1948), its discoverer. *Collimated light at the object plane* is its central thesis.

9.5 Travelling microscopes

Much metrology can be done by digital inspection of CCD images, sometimes by the investigator, sometimes by a computer program; and in this case it is the original optical instrument which must be calibrated by test measurements on known and well-established references. However, for many people this is a doctrine of perfection and for the most part we have to make do with the highly accurate and expensive common-user instruments which have often been around in the laboratory since pre-computer days. This applies in particular to travelling microscopes and plate-measuring machines, which stand carefully covered for dust protection and are only occasionally used.

A travelling microscope is generally used for measuring plates in spectroscopy and astronomy and comprises a compound microscope of moderate resolution mounted on a kinematic slide, with a calibrated screw for adjusting its position. While it is for the most part self-explanatory, there are a few points to be made about its use.

(1) There must be no parallax between the image and the eyepiece graticule. Adjust the cross-hairs first so that their image in the eye is at $-\infty$; then focus the object and check for parallax by small movements of the eye in the exit pupil.
(2) Choose a suitable graticule. Bear in mind that the eye is very sensitive to small inequalities. If the specimen is a spectrum with emission lines, a graticule with a pair of parallel lines, one on either side of the spectrum line, will give a more repeatable measurement than a single line or a pair of cross-hairs at 45° to the line. If the specimen is an astronomical field and a two-dimensional travelling microscope is used (a plate-measuring machine), a circular graticule will allow a stellar image to be centred more repeatably than a cross-hair, and furthermore with less discomfort to the operator.
(3) All mechanical screws should be assumed to have backlash. That is to say, there is a small 'dead' period when the screw is turned in the reverse direction. (Ignore the claims of 'anti-backlash screw' proponents!) Travelling microscopes typically have screws with a pitch of 1 mm and large drums at the end with 100 divisions and a vernier for interpolating tenths of a division. Measurements should always be made by turning the screw in the same direction, and the measurement should be repeated for the sake of consistency. If exactly the same reading is obtained each time, the measurement is not fine enough and a higher magnification should be used.

If extreme precision is sought, the measurements should be repeated with the plate in a different position in the stage. This is to check for the (unlikely) presence of a periodic error in the screw. Such periodic errors were once the bugbear of constructors of grating-ruling engines, but interferometric control of the ruling tip has eliminated this particular source of error.

10

Siderostats and coelostats

10.1 The celestial sphere

For the purposes of observational astronomy and celestial navigation, it is adequate to assume that the Sun, stars and planets all lie on the surface of a transparent sphere of 'immeasurably' large diameter with the Earth, or sometimes the observer, at its centre. The positions of the stars on this sphere are given by coordinates similar to those of latitude and longitude on Earth, but known generally as *declination* (latitude) and *celestial longitude*, measured westward from 'the first point of Aries'. This is the place on the sky where the Sun crosses the equator going northwards each (northern) Spring and was once in the constellation of Aries. (Because of the 'precession of the equinoxes' it is now in the constellation of Pisces.)

Formerly longitude was measured *eastward* from the first point of Aries. It was called 'right ascension' and was measured in hours and minutes rather than degrees. These hours were *sidereal* hours not solar hours, with the understanding that Earth rotates through 360° in 24 sidereal hours, or 23 hours 56 minutes of solar mean time. Sidereal clocks gain 4 minutes each day on ordinary clocks since the Earth rotates through approximately [1] 361° from one midday to the next.

10.2 Equatorial telescope mountings

The problem facing astronomers who wish to point a telescope at a particular fixed point in the sky or at the Sun, was solved by attaching the astronomical telescope to an axle parallel to Earth's polar axis and then setting the axle rotating at the sidereal or solar rate.

[1] Variable: it depends on the time of year.

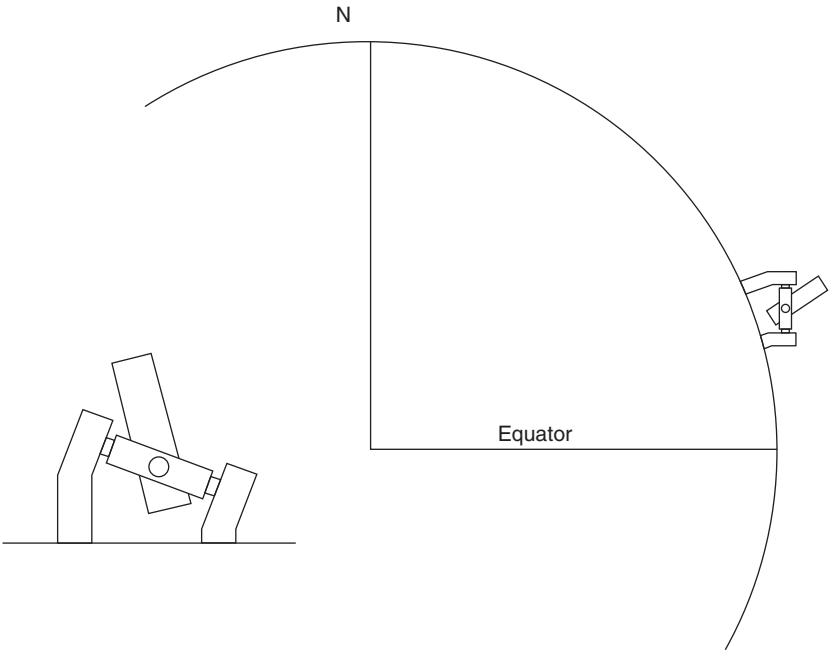

Figure 10.1 The principle of the equatorial mounting. The observatory in the diagram is at latitude 19° north. There are many versions of the equatorial mounting, designed to hold different types of telescope. All depend on rotation about a polar axis, an axle parallel to the axis of the Earth, turning at the sidereal rate while the telescope tube rotates about the *declination axis* allowing it to point indefinitely at a particular celestial object.

Various types of such 'equatorial' mounting have been devised, each with its own merits and demerits. However, the amount of engineering needed for an equatorial mounting is quite considerable and these days very large telescopes are generally mounted on so-called 'alt-azimuth' mountings, saving much weight and cost, and the 'pointing' is maintained with computer-driven motors. There is one awkward complication [2] with an alt-azimuth mounting: the image rotates as the night progresses and must be de-rotated by some optical device or by rotating the camera or spectrograph. These devices with their various internal and external reflections complicate the art of stellar polarimetry wonderfully.

[2] There is more than one in fact: the 'slew-rate' of the vertical axle is infinite if the celestial object passes through the zenith.

10.3 Coelostats

These are more properly described as *siderostats* as it is the celestial sphere that they are meant to track. The common name comes from the need to track the Sun at a solar eclipse for example, and the difference in tracking rate is about 1° a day or 2.5 seconds of arc per a minute.

10.3.1 Two-mirror coelostats

A common way to look at a fixed point in the sky is to use a mirror, rotating at the sidereal or solar rate, to direct the incoming chief-ray from an object, up or down a tube parallel to Earth's axis. A second mirror then directs this ray in any desired direction, usually horizontally in a laboratory where various instruments can be put in its way. This type of coelostat may be inconvenient if the object of interest is at a high opposite declination, because of the large angle between the mirror normal and the incident and reflected chief-rays, but is commonly done in solar observatories. Again there is a problem with image rotation.

10.3.2 The Lippmann coelostat

When an eclipse of the Sun is to be observed a similar problem must be solved and, furthermore, the solution must be portable, sometimes through rough country.[3] An elegant solution was found by Gabriel Lippmann[4] (1845–1921). This type of coelostat is now practically universal on eclipse expeditions. A large plane mirror is mounted in a cell which can be made to rotate about an axis parallel to the surface of the mirror. This axis of rotation is adjusted to be parallel to Earth's axis and the mirror is set to rotate at exactly half the sidereal (or solar) rate. The geometry of the method is elementary. The mirror normal necessarily points towards the celestial equator. By the laws of reflection, the incoming chief-ray from a celestial object, the mirror normal and the outgoing chief-ray all lie in the same plane, with equal angles between the rays and the mirror normal. If the mirror normal advances at half the sidereal rate the outgoing chief-ray will be constant in direction.

The observer is now at liberty to place his telescope anywhere, looking into the mirror obviously, provided that its optic axis is directed to some point on

[3] Sometimes by elephant, as in Sir Arthur Schuster's expedition to the 1875 Siam eclipse.

[4] Who incidentally received the 1908 Nobel prize for physics for his invention of a process of colour photography.

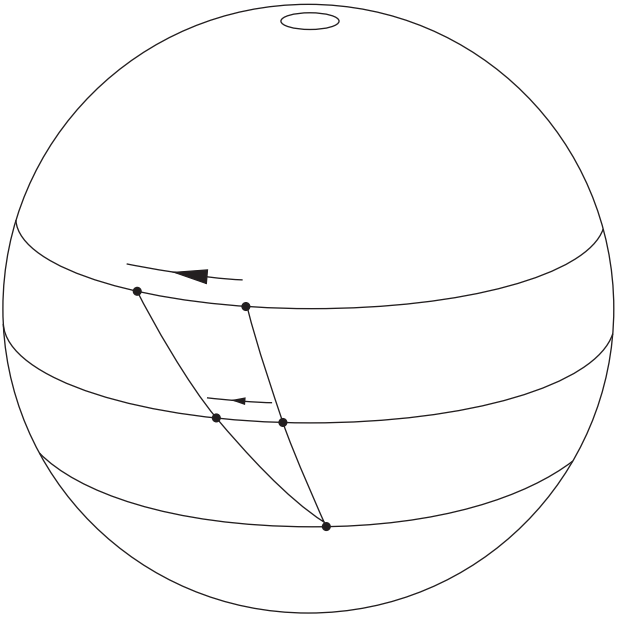

Figure 10.2 The celestial sphere, showing the point where the coelostat mirror-normal meets the celestial equator. The mirror rotates at half the sidereal rate, so that the reflected chief-ray of the object of interest remains stationary with respect to Earth's surface. The point on the equator to which the mirror-normal is directed is at the disposal of the observer, which means that the coelostat mirror can be rotated on its axle and clamped so that the reflected chief-ray is parallel to the observer's horizon. This allows the telescope to lie flat, looking into the coelostat mirror with its optic axis directed towards the point on the horizon where the object of interest will later set.

the latitude (declination) of the object of interest. The obvious thing to do then is to lie the telescope horizontally, facing towards the point on the horizon where the object will set later in the night (or later in the day if it is a solar eclipse). The coelostat mirror will then reflect into the telescope a point on that latitude, and by turning the coelostat on its axis to the appropriate place, the object of interest will appear in the telescope. If the mirror is clamped and the coelostat axis is turned at half the sidereal rate, the reflected object will appear stationary in the telescope.

An obvious corollary is that the telescope may be made to point towards the place where the object will rise. Which way is chosen depends on the position of the object in the sky when the observation is to be made. An afternoon

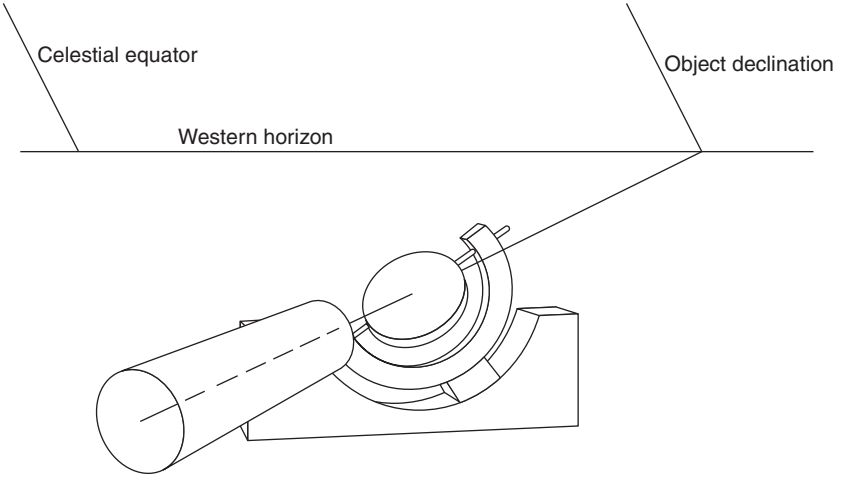

Figure 10.3 The Lipmann coelostat set up for an eclipse observation. The polar axis is adjustable for different latitudes and the turning mechanism (which is weight driven) can be adjusted to rotate at the solar or sidereal rate in either direction. The whole coelostat can be dismantled and packed in boxes which are (1) watertight so that they can be landed on desert islands through surf and (2) weigh less than 40 kg so that they can be carried by two people, a mule or a camel.

observation, for example, will require the latter choice, with the telescope pointing towards some point on the eastern horizon.

The advantage of the Lippmann coelostat is that the image does not rotate. The disadvantage is that it is not possible to observe objects at high declination. This is no problem for eclipse observation since the Sun's declination is never greater than $\pm 23^\circ\ 27'$. It has the further practical disadvantage that it is tedious in the extreme to adjust it so accurately that the image will be free from drift over a period of hours. A few seconds of arc or so of misalignment gives no problem in the 5 minutes of an eclipse observation, where the 'seeing' is no better than 10 seconds of arc, but it is unsatisfcatory for general observations unless it is installed and adjusted permanently.

A Lippmann coelostat is simple to construct and presents no great difficulty in initial alignment[5] at the observation site. It is not essential that the rotation axis lie exactly in the plane of the mirror, only that it be parallel to it.

[5] It is as well to make sure that the tracking motor is reversible when planning an expedition to the other side of the equator. Failure to ensure that the rotation direction is correct can lead to embarrassing moments and ribald comments from other observers at the eclipse site.

Spirit-levels and an inclinometer should be included as an aid to installation and alignment. [6]

[6] For people needing more information than is appropriate here, a useful and detailed account of coelostats in general is to be found in A. A. Mills, Heliostats, siderostats and coelostats – a review of practical instruments for astronomical applications. *J. Br. Astron. Assoc.* **95**, 89–99 (1985).

11

The detection and measurement of radiation

Which device is to be used to detect radiation depends very much on the kind of radiation. Heat, for example, the far infra-red that is, needs a bolometer; the UVOIR [1] region relies on the interaction of radiation with electric charge, and the far ultra-violet relies on fluorescence or the photoelectric effect.

11.1 The geometry of radiation measurement

Two things are fundamental.

(1) There is no power in a parallel beam: power depends on divergence into a finite solid angle.
(2) There is no power from a point source. The power radiated is proportional to the source area if the radiation is incoherent, and is proportional to the *square* of the source area if the radiation is coherent across the area.

The two useful ideas – a point source and a parallel beam – are, like light rays themselves, among the 'convenient fictions' of physics.

11.1.1 Extended sources

As we can see from the remarks above, all sources are extended to some degree. Even a single atom radiates a spectral line from an 'area' of about λ^2 because you cannot 'localize' the emitter to better than this, *using this radiation*. [2] Light sources then are distinguished by their *surface brightness*, the power emitted per unit area per unit solid angle. This quantity, multiplied by

[1] Ultra-violet, optical, infra-red.
[2] Light of a short enough wavelength is powerful enough to disrupt or destroy the source.

the étendue of an optical instrument, i.e. the entry-pupil area × the solid angle subtended by the detector, is the power received by the instrument.

11.1.2 Lambert's law

This states that the emission of radiation from a unit area of a surface varies as the cosine of the emitting angle. This is a purely geometric effect as the apparent area also varies with the cosine. The result is that from a Lambert radiator it is impossible to tell the tilt of the emitting surface with respect to the line of sight. A Lambert sphere looks like a disc. For the same reason it is impossible to tell how far away it is. If a distant surface has the same surface brightness as a close one, a detector with a given étendue will image a larger area of the more distant surface, thus cancelling out the inverse square law of radiation. It is the étendue of the instrument which determines the amount of radiation which it detects from an extended uniform source. Thus surface brightness is measured in watts per unit area per steradian, normal to the surface.

This last condition 'normal to the surface' is important because in Nature there are few Lambert radiators. The disc of the Sun for instance has only half the surface brightness near the limb that it has at the centre. A true spherical Lambert radiator would appear as a uniformly bright disc.

If the source is not extended enough to fill the étendue of the instrument, that is, if the image of the source is not big enough to fill the area of the detector, then the solid angle subtended by the source determines the amount of power measured. Ideally the optics should be designed, when the source is of small extent, so that its image just fills the detector area. You can do no better: you should do no less.

11.1.3 Point sources

As we saw above there are no strictly 'point' sources. The image of a star, for example, is not less than the size of its Airy disc at the detector. If there are aberrations (and there nearly always are) then the area over which it presents light at the detector is bigger than this, never smaller. The diffraction disc of a reflecting telescope is different from an Airy disc because of the central obstruction of the aperture and the consequent annular shape of the wavefronts before they are bought to a focus. In this case the central maximum is narrower than the Airy disc [3] but a greater fraction of the radiated power goes into the secondary annuli which surround the central peak. This effect is not trivial and

[3] Which has given rise to claims that such a telescope has a greater spatial resolution than an equivalent refractor.

the photometric aperture must be wide enough to include the first annulus if there is to be consistent photometry. When a single point source is being measured in astronomy, the 'seeing' blur [4] must also be considered, and so must the small sub-arc-second tremors of the telescope itself, not to mention its drive. All contribute to the noise on the signal: none can safely be ignored.

11.2 The detection of radiation

In practice, a radiation detector is a device for collecting and measuring the electric charges liberated from their constraints by the interaction of light with matter. This is not a comprehensive definition of course as it excludes bolometers which detect the heat generated by the absorption of radiant energy.

At least one Nobel Laureate, Willis Lamb, considered that the concept of a photon *as a particle* is unnecessary. [5] 'Photons', a quantum-mechanical concept, are unobservable. What we observe and measure are the consequences of the interaction of electromagnetic radiation with free or bound electrons in the solid, liquid, gaseous or plasma state. The electric charges thus released or otherwise stimulated can be collected, amplified and processed in various electronic and chemical ways (including the retina of the eye) to give the required information. The engaging question arises: are photons one of the convenient fictions of science?

Most likely to be found in an optics laboratory are detectors which depend on either photo-emission, photovoltaic effect or photoconductivity. In each case 'detection' amounts to the counting of photoelectrons released into the conduction band of a semiconductor, trapped in a p-type semiconductor or stimulated into the vacuum inside a photomultiplier. Photoelectrons can be stored temporarily in potential wells as in a CCD detector, to be regurgitated in sequence at a later time, microseconds later or years later. In a photomultiplier a single photoelectron emitted from the photocathode is accelerated by the voltage drop to the adjacent *dynode*, a grid held at about 120 volts positive with respect to the photocathode, where its energetic arrival stimulates secondary emission towards successive dynode stages. Each stage emits between three and four electrons per impact, and after 13 stages the charge reaching the anode is in the region of $10^7 e$, and enough charge is obtained for the pulse, a few nanoseconds long, to be amplified and counted in the same way as charge-pulses from nuclear events are counted.

[4] About 0.5 seconds of arc on a still, calm night, but 10 seconds of arc in daylight.
[5] W. E., Lamb, *Appl. Phys.* B **60**, 77–84 (1995).

In either case there is photon shot-noise [6] with noisepower as usual proportional to the squareroot of the signalpower [7] and *this is independent of the properties of the detector.*

11.3 The classification of radiation detectors

It is beyond the scope of this book to discuss details of the physics of the various detectors and we limit the discussion to the properties, merits and demerits of the various types of detector used. Different regions of the spectrum require different techniques.

11.4 The far infra-red

Radio wavelengths are detected by *coherent* detectors, rectifiers with local oscillators, which can detect and process information about the *phase* and *amplitude* of the incoming radiation. Coherent detectors cease to work [8] satisfactorily at wavelengths below a millimetre, and from here on down it is incoming *power* which is measured, usually by bolometers of one sort or another. As wavelengths grow shorter, other methods involving solid-state properties and semiconductor properties become more and more useful, and they have this in common: they are *analogue* devices which produce a current or a voltage proportional to the power they absorb. The signal must be digitized before being recorded. They also have an intrinsic random output, traditionally called 'noise', which depends on several things: on the sensitive area, on the operating temperature, on the a/c frequency of the signal, on the level of digitization, i.e. 8-bit, 16-bit and so on. Some of these noise sources are 'white', meaning that they produce equal amounts of power per unit fequency interval at all frequencies, and some are frequency dependent, the noise power decreasing at higher signal frequency. [9] Such detectors use a 'chopper', a rotating disc with alternate opaque and transparent sectors to provide alternating power from the source, with a phase-sensitive detector running at the chopping frequency to rectify the signal.

[6] Actually photoelectron shot-noise – but the effect is much the same.

[7] Here we mean that the RMS fluctuation in count from sample to sample is proportional to the square-root of the average count-per-sample, 'power' here being proportional to count-rate, that is, to electric current.

[8] At the time of writing, we do not pretend to predict the future in this sort of technology.

[9] Sometimes called 'pink noise'.

11.5 The middle infra-red

In the middle infra-red region, 10 μm to 50 μm there are several techniques available. Bolometers (strictly *microbolometers*) are insensitive to wavelength and measure just the thermal energy carried by the radiation. They use materials which have a large variation of some measurable property with temperature. Vanadium pentoxide, V_2O_5, is regarded as the material of choice because of its abnormally high thermal resistance coefficent. It is used uncooled. Microbolometers may be single detectors or pixel-arrays for thermal imaging in the 8–13 μm region where the atmosphere is transparent.

Other bolometric effects include thermal expansion (the Golay cell), the thermoelectric or Seebeck effect (in other words, thermocouples) and the pyroelectric effect in which heating or cooling a crystal induces a small *and temporary* voltage change across the ends. Bolometers generally are limited in their frequency response because of their thermal inertia and this may be a constraint on the design of an experiment.

Superconducting bolometers depend on the extraordinarily rapid change in electrical conductivity with temperature at the edge of the superconducting region. Temperature control must therefore be very precise and the surroundings must also be cooled (including the mirrors and lenses of the optical instrument providing the signal) to prevent contamination of the signal by thermal, i.e. black body, emission. Even from a polished surface this can be enough to drown a faint signal in thermal noise.

11.6 The UVOIR region

This acronym strictly applies to the visual region and the regions adjacent to it where radiation energies range from 1 eV to 10 eV, i.e. from $\lambda \sim 1.25$ μm to ∼0.125 μm or 1250 Å. This is the most interesting region for many scientists, where radiation interacts most strongly with matter, where molecules are stimulated and dissociated and atoms are ionized and recombine radiatively. This is the region where chemistry takes place, where the energies of chemical reactions are comparable with the energies emitted and absorbed by radiation.

11.6.1 Photoelectric effects

In this region photoelectric effects predominate, wherein the radiation excites electrons

(1) from a surface into the vacuum (photoelectric effect),
(2) from the valence band to the conduction band in a semiconductor (photoconductive effect),
(3) from the p-side to the n-side of a rectifying junction (photovoltaic effect).

There are several ways of using these effects.

(1) Each photo-induced emission event can be amplified and recorded as a pulse, as in a photomultiplier configured as a 'photon counter'.
(2) A succession of such events can be recorded as an electric current, rarely used now in instrumentation except in the superconducting mode.
(3) A succession of events can be stored and the accumulated charge subsequently amplified and measured, as in charge-coupled devices.

The efficiency of photoelectric devices is very high and in the photomultiplier for example, up to 40% of the incident radiated energy is absorbed and converted to a useable signal. As the efficiency is limited by thermal electron shot-noise rather than thermal noise, the noise is reduced drastically by cooling the photo-sensitive area of the detector. For many practical purposes a single-stage Peltier effect refrigerator is ample and will reduce the temperature of the sensitive area to 20 °C below ambient. However, for ultimate performance the device should be cooled, either to −80 °C by dry ice (solid CO_2) or to −180 °C by liquid nitrogen (LN_2). In a photomultiplier the random noise level is then 'a few' (i.e. less than 10) electron counts per second. In field conditions, for example at an eclipse, Peltier cooling is the obvious method of choice. [10] In a photomultiplier used for photo-current measurement, the EHT required at the cathode must be very strictly controlled since the current-multiplication varies as the nth power of the voltage, where n is the number of multiplication stages, typically 12–14. The EHT control is not so important in the pulse-counting mode unless the count-rate is so low that small pulses from spurious events (radioactive 1.33 MeV β-decay of potassium-40 in the glass wall causing Čerenkov radiation) are confused with genuine photocathode events, and then a pulse height discriminator must be used to eliminate pulses of lower height.

The practical limitation of photoelectric devices is that they lose sensitivity rapidly at longer wavelengths and, apart from some specially prepared photo-emitting surfaces such as indium gallium arsenide, are virtually useless at wavelengths longer than 1.2 μm. [11] They are sensitive – with quartz windows – in the ultra-violet but again lose sensitivity rapidly at wavelengths

[10] Although there are portable nitrogen liquefiers suitable for field work.
[11] Or 12,000 Å or 1200 nm or 1 eV (depending on the field of your endeavour!).

below 250 nm. For far ultra-violet meaurements, photocathode windows can be treated with a transparent fluorescent material, typically sodium salicylate, to augment their performance. A photomultiplier can conveniently be converted temporarily for use in the ultra-violet by coating the glass over the photocathode with a volatile solvent in which a heavy mineral grease [12] has been dissolved in dilute ratio. As an opaque reflective scattering screen on the other hand, sodium salicylate was best.

Photomultipliers are also subject to the following problems.

(1) Phosphorescence of the material itself (the glass envelope of a photomultiplier takes about an hour to recover from exposure to room light and it is worth painting the glass with a water-based colloidal graphite paint such as 'Aquadag' to counteract this).
(2) Local radioactivity, chiefly Potassium-40 in the glass envelope, and cosmic radiation, which can manifest itself either directly by ionizing the sensitive area or indirectly by inducing Čerenkov radiation in the surrounding transparent optical components.

Photoconductors in general use are chiefly lead sulphide and lead selenide, and are used, uncooled usually, for near infra-red detection. Lead sulphide covers the 1 μm to 2.9 μm region and lead selenide from 1.5 μm to 4.8 μm. Narrow band-gap semiconductors extend the range further. Indium antimonide, InSb, is sensitive from 1.0 μm to 6.5 μm and mercury cadmium telluride sensitivity extends as far as 25 μm.

11.6.2 CCD detectors

These are by far the most important of photodetectors for the visible and near ultra-violet region. The efficiency of a CCD detector is more than 60% at the time of writing, and the device is of such overwhelming importance in optical measurements that a separate chapter (Chapter 5) has been devoted to it and its idiosyncracies.

11.7 The far ultra-violet

Optical glasses are generally not transparent below 3000 Å, and often not below 4000 Å. Ultra-violet detectors therefore will be photoelectric detectors with quartz windows or, more likely, will rely upon fluorescence to convert

[12] In the author's experience as a graduate student, the 'apiezon' grease used in dismountable vacuum systems worked well: silicone vacuum grease did not.

the radiation to longer wavelengths. Quartz windows are satisfactory down to about 1800 Å and sapphire to 1600 Å, but at wavelengths below that, magnesium fluoride, calcium fluoride and lithium fluoride are the only known materials with adequate bulk transmission, i.e. from which millimetre thick optical windows can be made.

Windows are necessary at these wavelengths because oxygen begins to absorb strongly at 2300 Å and nitrogen at about 1800 Å, so that spectrometers working below these wavelengths are evacuated, usually to a pressure below 10^{-6} mmHg. The spectral region from 2000 Å to 100 Å is called the 'vacuum ultra-violet'. At 100 Å the region of soft X-rays begins.

LiF is transparent to about 1100 Å (the Balmer α-line of hydrogen at 1215.6 Å is transmitted at about 60%). For shorter wavelengths either fluorescent screens or *in vacuo* detectors such as electron multipliers are needed.

11.8 Types of spectra

There are broadly three kinds of spectra: line emission, line absorption and continuous. Emission sources can also be described as 'optically thin' and 'optically thick' and the spectra they emit depend strongly on the optical depth of the emitter. For example, a thin gas of sodium vapour, when stimulated electrically, will emit the D line doublet at wavelengths 5890 Å and 5896 Å with the intensity ratio of 2:1 predicted by the excited state multiplicity given by atomic theory. However, in an optically thick gas, absorption and emission conspire to increase the line widths and reduce the intensity ratio until, in a sodium street light [13] the intensity ratio of the D lines is nearly 1:1. This is because where emissivity is low, so is absorption. Taken to its logical conclusion, an infinitely deep cloud of sodium atoms will emit like a black body – all absorption and emission will have evened out and the temperature of the cloud alone will govern the net emission and absoprtion, and hence the shape of the spectrum.

Absorption spectra are the most difficult to measure accurately. Both the width and depth of an absorption line are affected by the scattered continuum light in the spectrograph and by its instrumental profile – the distortion that the instrument produces on a strictly monochromatic line. Consequently the concept of 'equivalent width' appears in the literature, measuring the amount of energy that has been removed by the absorpting medium between the source and the detector. This is obtained by measuring the area of the absorption because, unexpectedly, this is independent of the instrumental profile.

[13] Even a low pressure one.

Equivalent width is the width of a 'block' of the spectrum of full depth, with the same area as the measured line.

11.9 Black body radiation

A black body by definition is one which absorbs all the radiation which falls on it and which emits radiation according to the Stefan–Boltzmann law, and with a spectral distribution according to the Planck radiation law. These two fundamental laws are among the great achievements of nineteenth century physics. Together with Wien's displacement law they are the three important laws describing black body radiation.

11.9.1 Stefan's law

This is the law which connects temperature with the total radiated power from a surface. It is also known as the Stefan–Boltzmann law:

$$P = \sigma T^4 \tag{11.1}$$

where σ is *Stefan's constant* with the SI value 5.67×10^{-8}. The units of P are watts per square metre. The temperature is in degrees kelvin as usual.

This is the *total* radiation from a black body surface into the whole 2π solid angle above the surface. The *surface brightness*, S, the power per unit area, per unit solid angle *normal to the surface* on the other hand, is P/π, and *not* as one might expect, $P/2\pi$, because of the cosine factor in Lambert's law of radiation.

Again, this formula is an ideal. In practice it must be multiplied by the integral over all wavelengths of the emissivity function $e(\lambda)$ described below.

Typical values of emissivity for mirror coatings are 0.04 for newly evaporated aluminium, rising rapidly (in hours) to 0.12 as it oxidizes, and 0.05 for freshly evaporated silver, which it maintains provided it does not tarnish. Polished, uncoated, transparent glass surfaces have surprisingly high emissivity, in the region of 0.75 to 0.9 in the infra-red. This is because of absorption and emission in the interior as well as on the surface.

11.9.2 Wien's displacement law

This states that the wavelength of maximum intensity of the spectral distribution is inversely proportional to the absolute temperature. It can be stated briefly by the equation

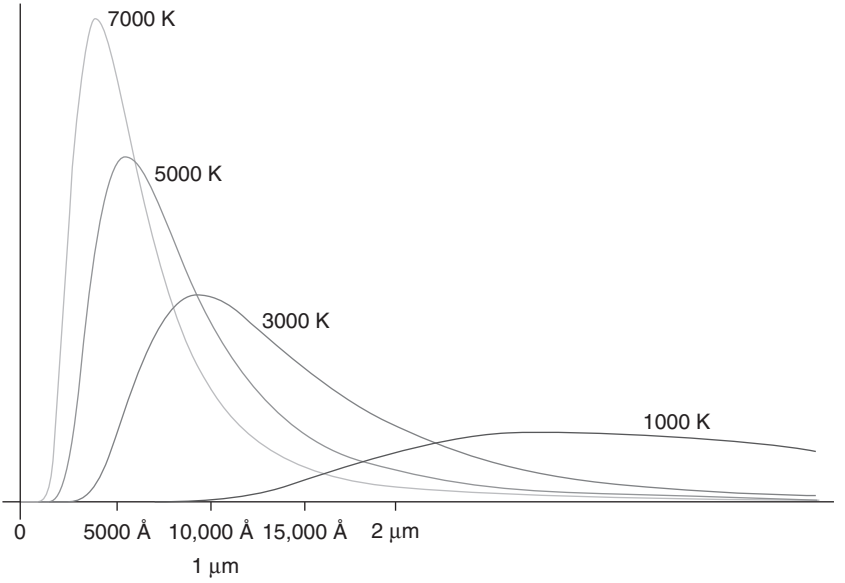

Figure 11.1 The Planck black body radiation curves for various temperatures. The curves are normalized to show equal total power. Notice that a hyperbola can be drawn through the various peak intensity points in accordance with Wien's displacement law. Notice also that the peak intensity for a black body at 5800 K is in the green, where the eye is most sensitive, also that the curve for a temperature of 1000 K has significant intensity only in the red part of the visible spectrum. The black body curve for a temperature of 20,000 K (not shown) will have its peak visible intensity in the blue/violet region, indicating why hot stars like Vega (α -Lyrae) appear blue in a telescope.

$$T.\lambda_{max} = 2.898 \times 10^{-3}\ \text{mK}$$

the units being metres and degrees kelvin.

11.9.3 The Planck radiation formulae

From the development of thermodynamics in the mid nineteenth century there were attempts to deduce the theory of radiation from thermodynamic principles. Early efforts by Balfour Stewart, Robert Kirchhoff, Wilhelm Wien and others led to the Rayleigh–Jeans law, a theory which worked well in the infra-red, but led to infinity in the ultra-violet, the so-called 'ultra-violet catastrophe'. The problem was resolved when a satisfactory explanation of the black body radiation spectrum was given by Max Planck (1858–1947) of the

University of Berlin, who simultaneously (and fortuitously) established the quantum theory of radiation.

The total power in watts, per unit area, per unit wavelength range emitted from the surface of a black body is given by:

$$I(\lambda) = \frac{2\pi hc^2}{\lambda^5} \frac{1}{\left(e^{\frac{hc}{\lambda kT}} - 1\right)} \tag{11.2}$$

and this must be divided by π to obtain the power per steradian emitted normal to the surface.

In the photoelectron counting regime, the corresponding photon emission rate is

$$N(\lambda) = \frac{2\pi c}{\lambda^4} \frac{1}{\left(e^{\frac{hc}{\lambda kT}} - 1\right)} \tag{11.3}$$

Everything is in SI units and due allowance must be made when converting to angstroms or nanometres of wavelength range and emitting areas in square millimetres.

11.9.4 Emissivity

The Sun approximates to a black body at a temperature of 5778 °C, a useful number when calibrating a spectrograph or photometer, provided allowance is made for atmospheric transmission. However, most hot objects do not radiate as black bodies because of varying *emissivity*, a quantity which depends on wavelength and accounts for the colours of objects. Polished metal surfaces, for instance, have a low emissivity and since the emissivity controls and describes absorption as well as emission, it is clear that the emissivity of even polished metal surfaces of gold and copper changes markedly with wavelength in the visible region.[14] Foliage, which is green in the visible, radiates like a black body in the infra-red. All surfaces have an emissivity function $e(\lambda)$ or $e(\nu)$ which multiplies the Planck radiation formula.

[14] There are persistent rumours that molten gold has a green cast, presumably complementary to the gold colour of reflected light, but this is not the author's opinion after evaporating gold to make infra-red mirrors.

12
Practicalities

12.1 Cleaning optical surfaces

You can clean a seriously stained or corroded surface by gentle polishing with a fine wet rag and a dab of rouge or cerirouge. This is a last resort of course, on a surface which would otherwise be discarded. Apart from this you should *never, never* touch a dirty optical surface with any solid other than a fine sable-hair artist's brush dipped in a suitable solvent such as analytical quality iso-propyl alcohol. The sweat on your fingers (which helps in fingerprinting) is acid enough to etch glass. There is *never* any reason to allow a cloth to touch an optical-quality glass surface, as it will almost certainly have abrasive silica particles embedded in it.

12.1.1 Camera lenses

Camera lenses are best kept with a permanent anti-ultra-violet filter in place. If it is necessary to clear dust, a blower-brush is the proper device. If the glass surface has stains or dried water-spots, then the solvent of choice is iso-propyl alcohol. Diethyl ether can also be used in a well-ventilated place and ethyl alcohol too, although it is not such a good solvent for lipids. Diethyl ether evaporates quickly and the consequent cooling of the surface may cause water-drop condensation in high humidity surroundings. Ethyl alcohol will dissolve water.

The basic rule is: don't touch and don't polish with a cloth.

12.1.2 Unmounted lenses

Apart from mounted lenses, optical parts almost always have some ground but not polished surfaces. These can be hand-held by the rough surfaces, preferably while wearing cotton or paper gloves. If a prism rests on a table, then one of the

ground surfaces should be at the bottom. A convex lens has to be surface-down unless it can rest on a hoop. It is safe on a freshly opened tissue from a box, folded and acting as a cushion. The enemy of scientific optical instruments is the scattered light which reduces contrast (thereby losing fine detail) and interferes with photometry, and it is dust particles, scratches and sleaks that do the scattering. Sleaks cannot be helped – they are part of the quality of the lens. Dust is removed by blower-brush (except on gratings). Scratches are what result from resting a convex glass face on a table.

Occasionally it is necessary, as an act of resurrection, to dismantle an old lens, taking its elements out one by one. Old nautical telescopes are particularly vulnerable to corrosion, and may have air-spaced achromatic objectives which need cleaning.

Method is important here: prepare beforehand a cylindrical block of wood or plastic with flat ends and with a diameter smaller than the smallest diameter in the lens body. Cover the end with a fresh cloth. Remove the retaining ring (usually a screw-thread but occasionally a retaining clip) and lower the lens on to the block so that the glass elements are pushed upwards out of the lens cell. Remove each one in turn and place it to one side *without turning it over* so that the orientation is not lost. *Method* is everything. Place each element as it emerges on a freshly unfolded tissue or some similar surface, side by side in its proper order and in the same orientation as it came from the lens barrel, with the cylindrical spacers which go between the elements all in their proper sequence, so that re-assembly restores the lens as it originally was. Whatever treatment is needed for each piece is performed separately and the piece restored. If any doubt is likely to arise about the orientation, then mark the edge by soft pencil with an arrow-mark so that it can be put back in its original place. Disassembly of any lens is not to be undertaken lightly and the proper preparation is paramount. An element replaced the wrong way round will ruin the resolution.

Cleaning is with one of the solvents mentioned above. Severe staining or 'blooming' of an old lens may be reduced, as mentioned above, by gentle rubbing with a wet cloth and a dab of jeweller's rouge or cerirouge, the two polishing materials used in lens making.

If a cemented achromat has to be separated, an electric hot-plate is needed. Mark the edge with opposed pencilled arrows *at the same point on the rim*. A low heat level is needed and the cement, Canada balsam most probably, will melt at about 300 °C but will probably soften enough before this to allow the elements to be slid apart. The two components can then be cleaned separately. Acetone will remove the cooled Canada balsam. To re-assemble the lens a lower heat is needed. The balsam is heated separately until it is free from

bubbles, and at the proper time a small drop, two or three cubic millimetres, is placed at the centre of a concave surface. Its mating surface is placed on top and the whole assembly squeezed hard together to secure the seal. A simple metal jig prepared beforehand will ensure the centration.

Needless to say, no attempt should be made on a modern zoom-lens. It is quite probable that special tools are needed anyway, and any work on the lens beyond brush-dusting should be left to the manufacturer's agent.

12.1.3 Plastic lenses

Lenses of optical research quality are commonly made from polymers of one sort or another and the rule still holds: don't touch! The manufacturer will have given instructions for cleaning and maintaining the outer surfaces and will say which solvents are allowable. Lenses, whether plastic or glass, which have had anti-reflection coatings applied will also require care and despite all the claims of manufacturers, are delicate and should be treated with the utmost care to avoid scratches. In general, if lenses are not in use they must be kept covered in an air-tight container.

12.1.4 Mirrors

Grubb–Parsons, the astronomical telescope makers of the twentieth century, used to advise that their mirror surfaces could be cleaned with 'anything that you would use to wash your face'. The reflecting surfaces were over-coated with silicon monoxide and were reasonably durable. Distilled water poured over a tilted surface removed dust,[1] insects and their pupae, pollen, bird-droppings and other detritus. Most optical instruments are not exposed to the elements like this however and a blower-brush will keep most mirror surfaces clean.

12.1.5 Diffraction gratings

It is only as a last resort that any attempt should be made to clean diffraction grating. If possible return them to the manufacturer for treatment. Otherwise dust particles (the most likely contaminant) can be dislodged by holding the grating with the surface nearly vertical and pouring a suitable liquid gently from a point just above the top edge. By 'suitable liquid' is meant anaesthetic-quality diethyl ether, a solvent which may remove some grease stains as well.

[1] Including fine, wind-blown Sahara sand in the Canary Islands observatories.

It must be done in the open air and *not* in the laboratory, except perhaps in a fume-cupboard, although even there, the thought of a spark from the motor of the extractor fan will concentrate the mind wonderfully. There is a latent problem, that if the local humidity is high, the cooling of the grating surface by the evaporation of the ether may result in the condensation of water. This evaporates in turn and may leave faint local marks.

Gratings not in use should either be covered as directed by the spectrometer manufacturer, or they should be dismounted and stored in an air-tight container.

12.2 Mounting optical parts

It may be convenient occasionally to convert a temporary arrangement to a permanent one, in which case fixed mountings will be needed for optical parts. Lenses and mirrors are generally mounted in a cell, a metallic cylinder which can be turned to shape in a simple workshop with a lathe. Glass and quartz expand thermally less than most metals, and likewise they contract less in cold surroundings. The fit should not be too fine therefore, in order to avoid any risk of compression and cracking of the glass in unusually cold surroundings (on going from 20 °C to 0 °C a gap of of 25 μm (0.001 inch) closes in a 75 mm diameter cell). The *centration* of a lens is generally about ±0.1 mm: that is to say, the optic axis and the centre of the rim of the lens coincide to this tolerance, so that a closer fit than this is of no optical advantage.

Both lenses and mirrors are held in place usually by one of two methods. A counter-cell is a ring with a threaded flange which is screwed into a corresponding thread in the cell, and tightened, not against the glass but against thin cork pads attached to the glass at the edge. If screw-cutting is not possible, then a face-plate may be used, this time with the flange pushed against the edge of the lens and held in place by three equispaced screws [2] (some people insist on six, for no obvious reason). Again there should be thin cork pads to prevent any sharp point-contact with the glass.

Prisms and large gratings usually stand on a plinth or table and are held in place by a bracket with a screw bearing down, via a soft, compressible mat, on the top edge. Traditionally lens cells were made of brass, but latterly a light alloy is preferred, for reasons not only of cost but of ease of fabrication. Most people buy their optical equipment: few have the facilities to make their own. For example, you buy achromatic lenses, you do not make them. If you need a special design, this is serious and expensive optics, only to be done following satisfactory ray-tracing with the manufacturer's list of tool radii beside you.

[2] In blind holes, not with retaining nuts.

When incorporating existing lenses in a new instrument, be aware that astronomical objective achromats were sometimes computed for correction in the blue and near ultra-violet rather than blue and green or green and red. They hark back to the days when photographic emulsions were insensitive to green, yellow and red light. They are generally known as 'astrographic objectives'.

Finally, as a general piece of philosophical advice: don't waste time and money buying or building a fancier instrument than you need, in the hope that it may come in handy for later, more elaborate research. It won't! It will merely end up gathering dust in the store room while a successful research project is extended with a bigger research grant and new, updated, appropriate equipment. The best you can hope for is to bequeath it to the teaching laboratory or the nearest high-school science faculty.

12.2.1 Cross-hairs and fiducial marks

Spider silk cross-hairs are still in use. They distort and break from time to time and need replacing. Any sort of animal hair is much too coarse, typically 25 μm in diameter. Spider silk is about 1 or 2 μm in diameter but is not always – or even usually – available. Stretching a spot of adhesive, especially old-time seccotine, a fish-collagen based protein glue, may give a thin enough thread if quartz fibre is not available.

Manipulating very fine fibre is done by attaching it to a card cut to the shape of a miniature horse-shoe, with the fibre across the open jaws and secured by a dab of glue or paste. This device, lowered gently on to the rim of the holder of the cross-hairs, will locate the fibre accurately while two spots of glue secure it in place.

12.3 Optical filters

Optical filters, like electrical filters, can be broadly divided into low pass, band-pass and high pass, but referring to optical wavelengths rather than frequencies. Some depend on natural or artificial dyestuffs and the 'stained glasses' of cathedral windows are glasses containing various metallic cations in solution. These have the particular advantage that they do not fade with time. This is not commonly true of dye filters, especially those which absorb the high energy of blue and ultra-violet light. Red dyes especially tend to fade and must be carefully selected. Red-stained glass, on the other hand, may admit blue or violet light enough to be a nuisance.

12.3.1 High pass filters

These are particularly useful in the near infra-red. They are to be found among the amorphous semiconductors which are transprent to radiation energies below the band-gap energy. The absorption begins at what is usually called the 'Urbach edge', and selenium for example transmits just on the edge of the visible region. A simple near infra-red high pass filter can be home-made by sandwiching a thin film of molten black [3] selenium between two glass plates. If it is thin enough it will just transmit enough ($<0.1\%$) red light to be seen.

High pass filters in general are not difficult to make with sharp edges, and the familiar orange and yellow filters of the photographer are typical examples. Low pass filters on the other hand are not easily contrived. Manufacturers provide graphs of their transmission characteristics.

12.3.2 Interference filters

These are Fabry–Pérot étalons with a very small gap, usually only half a wavelength thick, so that they work in first order. As with spectroscopic étalons, the reflecting surfaces are sandwiches: quarter-wavelength thick layers alternately of high and low refractive index material, typically zinc sulphide and cryolite. Seven layers is the normal complement and a half-wavelength thick layer of cryolite at the centre of the stack is the Fabry–Pérot 'gap' which determines the pass-band wavelength. These filters have a pass-band typically about 10 Å wide and a peak transmission in the region of 60–80%. As such they are useful order-sorters for Fabry–Pérot spectrographs, but must be used with care: the pass-band changes with angle of incidence, moving to a shorter wavelength and broadening as the filter is tilted. [4] Generally they are made to order for a particular investigation, most often with the pass-band centred on an astrophysically interesting line.

12.3.3 Notch filters

These are so-called because they remove a certain band of the spectrum. They may be of didymium glass or they may be a species of interference filter, manufactured by the same sort of vacuum evaporation process as interference filters. These are complementary to interference filters in having a narrow stop-band rather than a narrow pass-band. Again their use is in astronomy, where they can

[3] Selenium has two allotropes; black selenium melts at 220 °C.

[4] The change of pass-band with tilt angle allows them to be used as crude low resolution spectrometers.

be manufactured to prevent sodium or mercury street lighting from interfering with observations.

Didymium glass filters on the other hand are made of glass containing the rare earth elements praseodymium and neodymium, and they absorb a wide band in the yellow/orange region of the spectrum. Their chief laboratory application is to remove the yellow light in the spectrum of a mercury lamp, leaving the green mercury resonance line at 546.1 nm as a practically monochromatic light source for inspection interferometry.

It is less well known that an aqueous solution of a salt of europium has several narrow absorption lines in the blue and near ultra-violet parts of the spectrum and that these can be used as passive wavelength standards in a liquid absorption-cell.

12.3.4 Christiansen filters

These comprise a mixture of two finely divided solids or a solid and a liquid which are transparent, with refractive indices identical at one wavelength but with different dispersions. At the matching wavelength they are simple glass blocks, but elsewhere the incident light is scattered and they are opaque but translucent. The pass-band is temperature dependent.

Alternatively, a mixture of different crystalline materials in fine powder form, pressed or sintered together, has similar properties, particularly in the infra-red.

12.3.5 Dichroic filters

These are filters which are simultaneously high pass and low pass filters, and reflect the light that they do not transmit. Their general use is in colour photography and in particular colour television and cinematography where, depending on the wavelength cutoff point, they will send light in different directions to different CCD detectors. This obviates the need for a Bayer mask and so improves the picture resolution. They also have a place as infra-red reflectors in slide-projectors, allowing a brighter light to pass without the accompanying heat.

12.3.6 Order-sorters

These are filters which suppress unwanted orders of interference (in étalons) or diffraction (in gratings). Gratings are apt to show lines coming from light a half or a third of the wanted wavelength, and high pass filters are usually

required. Glasses are easily found which are opaque to ultra-violet light and these give adequate protection for spectroscopy in the visible region. The above-mentioned high pass filters perform the same office for spectroscopy in the near infra-red. Some caution though, is needed. The visible region occupies less than an octave of the spectrum. In the infra-red one is apt to require spectra covering, for example, from 2 μm to 4.5 μm or 8 μm to 13 μm – the atmospheric 'windows'. The latter offers no problem, but the former extends over more than one octave, so that a line at 2 μm will appear again at 4 μm. This makes a high pass filter highly desirable.

Appendix A

Gaussian optics

A.1 The purpose

To lay out a centred optical system of lenses and mirrors on a real or virtual drawing board, requires simple geometrical constructions, and Gaussian optics provides the rules for making them. Rays can be traced from an object point to its conjugate image point, positions adjusted and apertures decided to allow all necessary ray bundles to pass unobstructed and unvignetted through to the image.

A.2 The basic assumptions

We assume from the start that all angles and all apertures are small enough that the approximations $\sin\theta = \tan\theta = \theta$ and $\cos\theta = 1$ can be made. These assumptions are sufficient to allow the main layout of an optical system to be done graphically. When this basic layout is satisfactory, Seidel theory (described in Appendix B) can be invoked, and vertex positions, radii of curvature, stop positions etc. can be adjusted to minimize aberration, and computer ray-tracing can begin.

A.3 Refraction at a curved optical surface

The law connecting the position of an object and the position of its image after refraction comes from elementary geometry. From Euclid we have (Figure A.1)

$$\alpha + i = \gamma \qquad \beta + r = \gamma$$

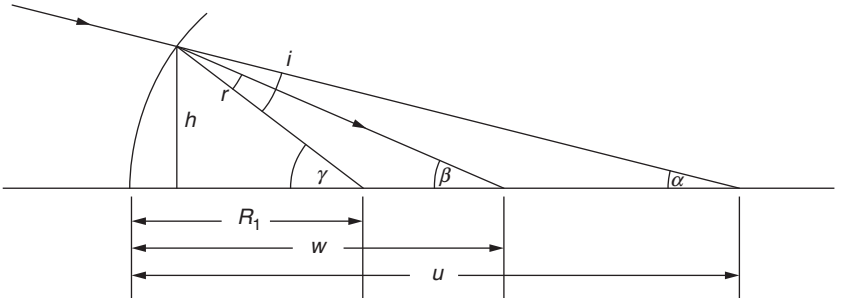

Figure A.1 Refraction at a curved surface. The rules of Gaussian optics apply: sines and tangents are reduced to the angle itself. Transverse dimensions such as h are considered infinitesimal.

and from the Gaussian approximation,

$$\alpha = h/u \quad \beta = h/w \quad \gamma = h/R_1 \quad i = nr$$

where u is the 'object distance', w the 'image distance' and R_1 the radius of curvature of the surface.

From this we find:

$$\frac{n}{w} - \frac{1}{u} = \frac{n-1}{R_1} \tag{A.1}$$

For the second surface similarly

$$\eta + \theta = j \quad \eta + \phi = k$$

where j is the angle of incidence *from inside the glass* and k the angle of refraction into air. $k = nj$ from the law of sines:

$$\theta = h'/w \quad \phi = h'/v \quad \eta = h'/(-R_2)$$

The minus sign before R_2 is because the centre of R_2 is to the left of the vertex and R_2 is therefore a negative quantity. [1]

Likewise we find for the second surface:

$$\frac{1}{v} - \frac{n}{w} = \frac{n-1}{-R_2} \tag{A.2}$$

Adding equations (A.1) and (A.2) we find:

$$\frac{1}{v} - \frac{1}{u} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

[1] If in doubt about this, remember that if both radii were the same the lens would be a meniscus and would not refract.

If the incoming ray is parallel to the optic axis, that is if $1/u = 0$, then the ray crosses the optic axis at the focus of the lens, and v is the *focal length* of the lens, f.

Opticians and ophthalmologists, and sometimes optical designers, use the reciprocal of the focal length, $1/f = P$, the *power* of the lens. P is usually measured in *dioptres*, the reciprocal of 1 metre, so that for example, a 3 dioptre lens has a focal length of 33.3 cm. It is also common practice among optical designers to use *curvatures*, the reciprocal of the lens radii, so that the formula for the power of a lens becomes:

$$P = (n - 1)\Delta c \tag{A.3}$$

where Δc is the difference between the first and second curvatures. This incidentally says more than appears. It is the *difference* in curvature which matters, not the curvatures themselves. So a lens may be biconvex, plano-convex or meniscus, always with the same focal length but with different 'bendings', and these bendings may have a significant effect on the quality of the image finally produced.

A.3.1 Images

If the image formed by a lens can be demonstrated on a screen it is said to be *real*. If, on the other hand, it is apparent only by looking at it through the lens, it is *virtual*. Single negative lenses, i.e. with concave surfaces and negative focal lengths, will always give virtual images. However, lens *combinations* may have negative focal lengths and yet give real images. Real images tend to be inverted with respect to their objects and virtual images tend to be upright. This is not always true of course for complicated systems of lenses, but is a useful guide for simple systems.

A.3.2 Application of the lens formula

Since there is sometimes some confusion it is worth giving a few examples of elementary calculations involving lenses. A useful precept is that the angle through which a ray is refracted depends only on the height of incidence and the focal length of the lens:

$$\theta = h/f$$

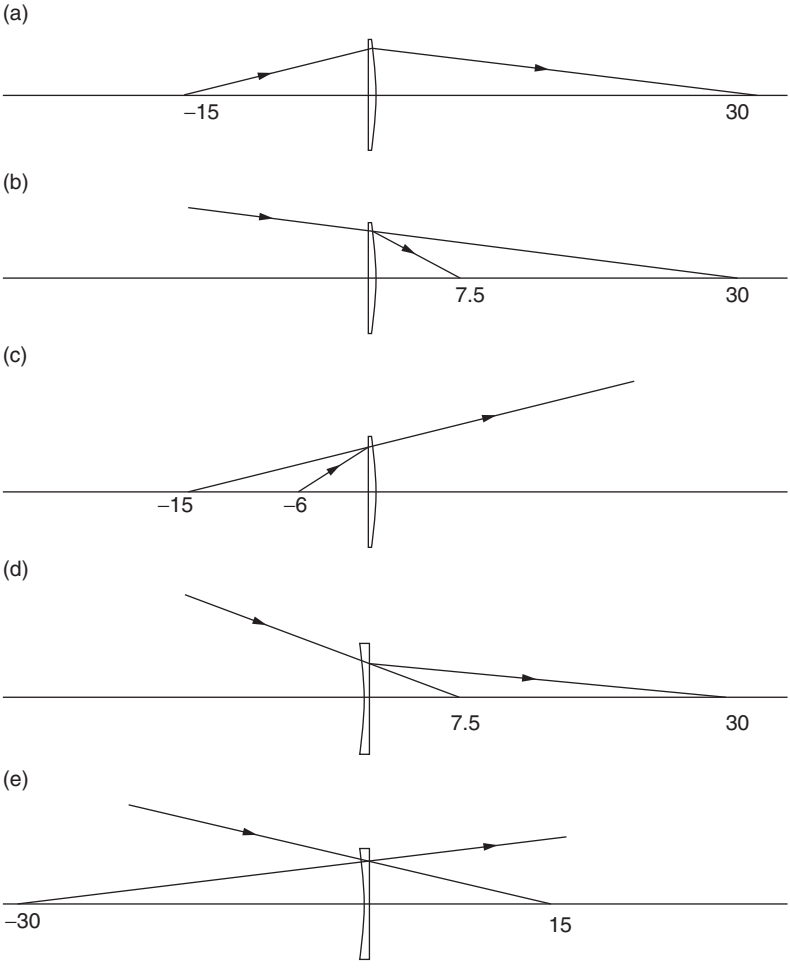

Figure A.2 Examples of refraction through thin lenses of focal length 10 (a, b, and c) or −10 (d and e). Diagrams such as this express the theory more succinctly than the algebraic formulae and can be used to check the formulae themselves.

and this can be used to examine the passage of light through two lenses of focal lengths f_1 and f_2 separated by a distance d:

$$\theta_1 = h_1/f_1$$

$$h_2 = h_1 - d\theta_1$$

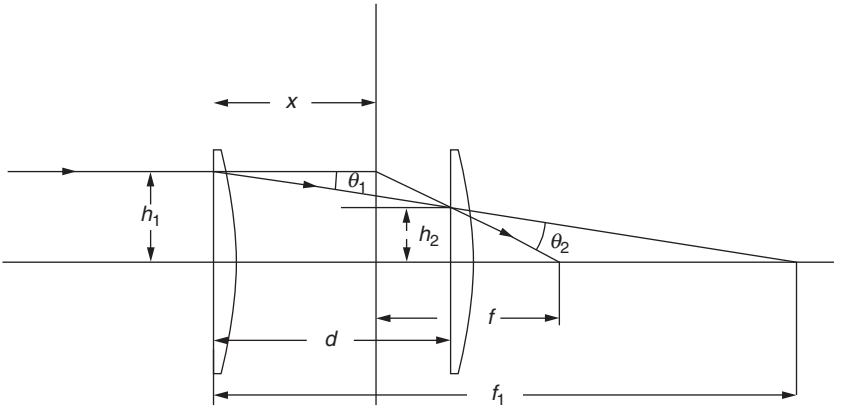

Figure A.3 The geometry of a ray passing through two lenses. As before, angles and transverse dimensions are infinitesimal. The position of the single equivalent lens is a distance x to the right of the first element.

$$\theta_2 = h_2/f_2$$

$$\theta_1 + \theta_2 = h_1/F$$

where F is the focal length of the combination.

Now replace the θ_1 and θ_2 to give:

$$\begin{aligned} h_1/F &= h_1/f_1 + (h_1 - d\theta_1)/f_2 \\ &= h_1/f_1 + h_1/f_2 - dh_1/f_1 f_2 \end{aligned}$$

If we discard the h_1 we get the two-lens formula:

$$\frac{1}{F} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2}$$

The focal length of the combination is

$$F = \left(\frac{f_1 f_2}{f_1 + f_2 - d} \right)$$

This all well and good but it does not tell us where the focal length is to be measured *from*. A few lines of geometry, however, will show that the two lenses could be replaced by a single equivalent lens of the same focal length as the

combination, provided it is placed a distance x to the right of the first lens, where x is given by:

$$x = \frac{d(f_1 - d)}{f_1 + f_2 - d}$$

A similar expression determines the position of the single equivalent lens to the left of the second component if light is coming from the right.

A.3.3 Principal points and planes, focal planes, nodal points – the cardinal points

This placing of a single equivalent lens leads to two useful ideas, When first laying out the optical design of a system it is well to bear in mind that for *any* combination of lenses and mirrors (it is not restricted to two) there are six planes which determine the refraction. These govern the rules for laying out the ray diagram. Where they cross the optic axis are the six *cardinal points* of the system. The *principal planes* are 'notional' planes, the places where you would put a single equivalent lens to get the same image of an object at $-\infty$. They cross the optic axis at the *principal points*. The *nodal planes* likewise are notional and for most purposes are coincident with the principal planes. They differ if the refractive index on the object side is different from that on the image side as, for example, in a lens which is designed to work under water or an oil-immersion objective in a microscope. The *focal planes* are real planes, where the image of an object at $\pm\infty$ is formed.

Warning! They are plane only in the Gaussian approximation. When apertures are large their positions move with refraction angle, and they are usually curved surfaces (although making the focal planes flat is one of the goals of a camera lens designer).

The rules for drawing rays are now as follows.

(1) A ray, parallel to the optic axis and coming from the left, arrives at the first principal plane P_1 at a height h_1. It emerges from the system from the second principal plane at the same height h_1 and crosses the axis at the second focal plane.
(2) A ray, crossing the optic axis at the first focal point, arrives at the first principal plane at a height h_2 and emerges from the second principal plane at the same height h_2 and parallel to the optic axis.
(3) A ray inclined at an angle θ, which crosses the optic axis at the first nodal point, emerges from the second nodal point *at the same angle θ to the optic axis.*

If the initial and final refractive indices are the same (which they usually are unless the lens is designed to work under water or is an oil-immersion microscope objective) **the principal and nodal points and planes are the same.**

Any two of these three conditions are enough to allow the ray diagram of a system to be constructed. There is no need to consider the various refractions that take place inside the lens system: the principal, nodal and focal points and planes alone define the ray geometry. In air the two focal lengths are the same, and the distance from each principal plane to its focal plane is the focal length. For a single thin lens the two principal planes coincide but the same rules apply for constructing rays. A ray crossing the optic axis at the lens is undeviated. A single thick lens has two principal planes, albeit very close together.

The precept mentioned above, regarding the height of incidence and deflection of a ray, has two other useful corollaries.

(1) A lens placed at the image formed by another lens has no effect on the image. Even if the second lens is of poor quality, the image remains intact. Even a ground glass screen placed there will not distort the image.
(2) If two identical positive lenses are separated by twice their focal length, the object-to-image distance is independent of the object distance from the first lens.

A.4 The achromatic doublet

This is a simple exercise in lens design. We choose two of glass types of different refractive index and different *partial dispersions* and combine them in such a way that the focal length of the combination is the same for two different wavelengths. It is to be a *cemented doublet*, so that the first curvature of the second element is the same as the second curvature of the first element. The difference is that one is convex and the other concave.

From the glass maker's catalogue we get the refractive index of each glass at the D Fraunhofer line, at 589.3 nm, and we also get the *partial dispersions* or Abbé v-values. These are important numbers which determine the dispersive power of the glass and are defined by

$$v = \frac{n_D - 1}{n_F - n_C}$$

where the n are the refractive indices at the Fraunhofer C ($\lambda = 656.3$ nm), D ($\lambda = 589.3$ nm), and F ($\lambda = 486.1$ nm) lines. For a typical crown glass, Schott BK7, $n_D = 1.5167$, $v = 64.2$ and for a dense flint-glass with high

dispersion, Schott SF11, $n_D = 1.7844$, $v = 25.6$. Paradoxically perhaps, the higher dispersive glass has the lower v-value.

Assume that the power required of the lens is P. As the two elements are in contact the power is the sum of the two individual powers. From equation (A.3) we have

$$\begin{aligned} P = P_1 + P_2 &= (n_1 - 1)C_1 + (n_2 - 1)C_2 \\ &= (n_1' - 1)/C_1 + (n_2' - 1)C_2 \end{aligned}$$

which is the condition that the lens combination shall have the same power at two different indices of refraction.

From this it follows that

$$\Delta n_1 C_1 = -\Delta n_2 C_2$$

and then P is given by

$$\begin{aligned} P &= (n_1 - 1)C_1 - \frac{\Delta n_1}{\Delta n_2}(n_2 - 1)C_1 \\ &\sim \Delta n_1 C_1 (v_1 - v_2) \end{aligned}$$

where the approximation sign implies that an intermediate wavelength (Fraunhofer line D) has replaced n_1 or n_2 in the v-formulae.

It is then easy to show that

$$P_1 = \left[\frac{v_1}{v_1 - v_2}\right] P \qquad P_2 = -\left[\frac{v_2}{v_1 - v_2}\right] P$$

Although the two elements are in contact, it does not mean that they are necessarily cemented. A close air-spaced pair of elements with different curvatures will still be achromatic, and the individual elements can be 'bent' to correct spherical aberration. The achromatic requirement only specifies the powers, ΔC_1 and ΔC_2 so that there are two more degrees of freedom, the two first curvatures of the elements, which can be used to correct geometrical aberrations.

A.5 Gaussian beams

A Gaussian beam is essentially what comes out of a gas laser. It is a beam of light with quasi-plane wavefronts which are coherent over a distance of about 1 mm, and it diverges at two or three minutes of arc. The amplitude of the field is assumed to have a Gaussian distribution, which is to say that (using

cylindrical polar coordinates r, θ, z with z along the optic axis) the electric field amplitude varies according to the Gaussian function:

$$A(r) = A(0)e^{-r^2/a^2}$$

It is a basic theorem in diffraction theory (see Appendix C) that when a coherent [2] wavefront is diffracted at an aperture, the amplitude of the diffracted beam is the Fourier transform of the amplitude across the diffracting aperture. It is also a theorem in Fourier theory that the transform of a Gaussian function is another Gaussian function. It follows then that the radial distribution of amplitude in a Gaussian beam is unchanged as it propagates, with only the beam diameter increasing or decreasing with distance and with passage through various optical elements. In practice, this is only approximately true as the Gaussian description is only an approximation, bounded as the beam is by its finite diameter.

Gaussian beams have considerable application in technology, and the optical components associated with them can be of the simplest kind, there being neither field nor aperture of any considerable extent. Likewise, as laser light is monochromatic, there need be no correction for chromatic aberration. They can be focused with simple non-spherical lenses to diameters of 0.5 μm or less, for example for the burning and scanning of DVDs.

A.6 The ray transfer matrix

This an alternative method of tracing a ray through a centred system at the level of basic Gaussian optics. If the height of a ray above the optic axis and its inclination are known, the effect of its passage through an optical element can be represented by a simple 2×2 matrix operation. The ray is defined initially by a column matrix:

$$\begin{bmatrix} h_1 \\ \theta_1 \end{bmatrix}$$

The ray is assumed to 'enter' a notional *input plane* [3] perpendicular to the optic axis at a height h_1 and at an inclination θ_1, and to 'emerge' at the *output plane* at a height h_2 and inclination θ_2. The two planes may be separated by a distance d, the distance between two optical vertices, but where the matrix

[2] Defined and described in Chapter 4.
[3] Not to be confused with the principal plane of Gaussian optics.

describes the passage of a ray at an optical surface or a thin lens, they are regarded as coincident.

As a formal mathematical statement the ray transfer equation is:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} h_1 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} h_2 \\ \theta_2 \end{bmatrix}$$

and the terms A, B, C, D in the square matrix depend on the particular transaction that is being described.

There are two advantages of this method of optical layout design. The first is that when there are many surfaces, refracting or reflecting, with varying refractive indices or vertex distances, the whole system can be represented by a single matrix, the matrix product – in the proper order – of the individual matrices representing the successive surfaces and separations. The second, and probably the more important, is in the analysis of optical resonators, devices which can store monochromatic wave-trains which complete circuits, sometimes with the resonant oscillation sustained by one or more optical amplifiers. The Sagnac interferometer is an example of an optical resonant cavity, and if a gas laser is included in the circuit it becomes a *ring laser*. The Fabry–Pérot étalon is another example, and when the gap between the reflectors is filled with a medium which can provide stimulated emission, it becomes a gas laser.

The technique of analysis by the ray transfer matrix method is best demonstrated by familiar examples. The 2×2 matrix

$$\begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}$$

pre-multiplying the ray matrix, extends the input ray a distance d to the right, changing the height of its end-point from h_1 to h_2 but not its inclination, thus:

$$\begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} h_1 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} h_2 \\ \theta_2 \end{bmatrix}$$

Which yields the two equations:

$$h_2 = h_1 + d\theta_1$$
$$\theta_2 = \theta_1$$

The matrix

$$\begin{bmatrix} 1 & 0 \\ \left(\frac{1-n}{Rn}\right) & \frac{1}{n} \end{bmatrix}$$

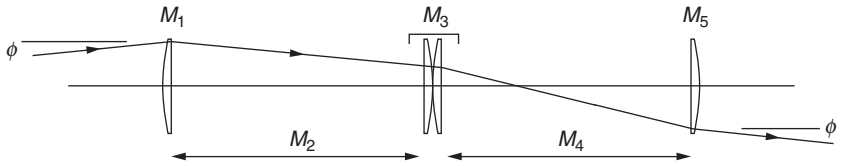

Figure A.4 The lenses of the relay section in the text. In common practice, all such diagrams, whether of lenses or mirrors, are laid out as if they were lenses. This avoids complicated ray-diagrams with overlapping rays travelling in both directions and with stop-positions which are confusing and not easily read.

passes a ray from air ($n = 1$) through a surface of radius R bounding a medium of refractive index n. Thus:

$$\begin{bmatrix} 1 & 0 \\ \left(\frac{1-n}{Rn}\right) & \frac{1}{n} \end{bmatrix} \cdot \begin{bmatrix} h_1 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} h_1 \\ \frac{(1-n)h_1}{Rn} + \frac{\theta_1}{n} \end{bmatrix}$$

which does not extend the ray, and changes only its inclination. Similarly the matrix:

$$\begin{bmatrix} 1 & 0 \\ \frac{(n-1)}{R} & n \end{bmatrix}$$

restores the ray to air. The two matrices multiplied together, with radii R_1 and R_2, give the matrix which describes the passage of a ray through a thin lens:

$$\begin{bmatrix} 1 & 0 \\ (n-1)\left(\frac{1}{R_2} - \frac{1}{R_1}\right) & 1 \end{bmatrix}$$

which can more succinctly be written

$$\begin{bmatrix} 1, & 0 \\ -\frac{1}{f}, & 1 \end{bmatrix}$$

and this, when post-multiplied by the column matrix containing h and θ, shows accord with the statement earlier, that the deviation of a ray by a thin lens depends only on its height of incidence and the focal length of the lens.

If the displacement matrix is interposed between the two refraction matrices above, the behaviour of a ray passing through a thick lens can be deduced. We leave this as an exercise for the reader in elementary matrix multiplication. Notice that this matrix gives the height and inclination of the ray as it leaves the second surface. It says nothing about the further progress of the ray.

A simple and instructive exercise is to trace a ray through a relay section, as might be used for example in an erecting telescope (Figure A.4). There are five

matrices needed: the field lens, the interval, the relay lens, the second interval and the collimator,

$$\begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & f \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ -\frac{2}{f} & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & f \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix}$$

and their product is the single matrix

$$\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

showing that all the rays from the original object at $-\infty$ have been transferred, inverted, to the next stage of the relay.

A.6.1 The optical resonant cavity

The gas laser began as a Fabry–Perot étalon with a small aperture, a large gap, containing a mixture of helium and neon which were selected to provide coherent amplification of one spectrum line. At first it was found extremely difficult to sustain resonance and the solution was to make one or both of the Fabry–Pérot reflectors slightly concave. The result was to provoke a number of resonant modes which emerged in several slightly different directions as Gaussian beams. The single mode gas laser evolved from this, and other types of resonant cavity followed, in particular the Sagnac interferometer with one or more HeNe lasers in the arms.

These devices, from the point of view of a Gaussian optics layout (where as usual a reflecting mirror is represented by a thin lens of the same focal length), were periscopes in which light was confined within the tube and relayed from one section to the next. The ray transfer matrix method showed the conditions under which this confinement would happen. The basic optical element, which the light is to traverse repeatedly, is the concave reflector plus the distance to the next reflector. There are four 2×2 matrices which represent one cycle of the resonant cavity. In sequence (from the right) they are:

$$\begin{bmatrix} 1 & 0 \\ -\frac{2}{R_2} & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ -\frac{2}{R_1} & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix}$$

and bearing in mind that the focal length of a concave mirror of radius R is $R/2$, the product of these is the $ABCD$ matrix of the cavity.

In order, they describe

(1) transfer to the reflector of radius R_1,
(2) reflection,

(3) transfer to reflector of radius R_2,
(4) reflection.

In short the cycle can be written:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} h \\ \theta \end{bmatrix} = \begin{bmatrix} \alpha h \\ \alpha\theta \end{bmatrix} \tag{A.4}$$

and in order that the ray be contained within the resonator cavity, α must lie between 1 and -1.

This is the *characteristic equation* of the resonator and is solved in the usual way by requiring the determinant to vanish:

$$\begin{vmatrix} A-\alpha & B \\ C & D-\alpha \end{vmatrix} = 0$$

and then α is a root of the quadratic equation

$$\alpha^2 - (A+D)\alpha + (AD - BC) = 0$$

In general α may not be real, but the modulus of a complex root of a quadratic equation is the square-root of the constant coefficient which in this instance is ± 1. The real part of the root must be less than or equal to this, so that

$$\left| \frac{A+D}{2} \right| \leq 1 \tag{A.5}$$

which results in the condition:

$$0 \leq \left(1 - \frac{d}{R_1}\right)\left(1 - \frac{d}{R_2}\right) \leq 1$$

This tells us the allowable limits on the radii of the end-mirrors for a given separation d. In the early HeNe lasers both radii were infinite – at the limit allowed for containment – and it was learned at an early stage that it was unwise to approach either limit. In practice one plane mirror situated at or near the focus of the concave mirror, made resonance relatively easy to acquire.

It is worth adding that this was discovered empirically, not as a consequence of this analysis. Having the two concave reflectors concentric probably makes resonance equally difficult to achieve.

Appendix B

Optical aberrations

B.1 The limits of Gaussian optics

The initial layout of an optical instrument is usually made using simple Gaussian optics, where all apertures and pupils are infinitesimal and all angles are considered small, so that the approximations $\sin\theta = \tan\theta = \theta$ and $\cos\theta = 1$ are adequate.

Geometrical optics, however, is a non-linear discipline and ray-tracing, using the proper values of sines and cosines, quickly reveals that rays do not arrive in the expected places to make sharp images. The quality of an image can be tested by modifying the Gaussian optics to the approximations that $\sin\theta = \theta - \theta^3/6$, $\cos\theta = 1$ and $\tan\theta = \theta + \theta^3/3$ and adjustments can then be made to the positions, curvatures and refractive indices of the various elements of a system to produce better images at the focal surface. Except in rare instances, perfection is not to be expected, and the art of the optical designer is to adjust the values to produce an acceptable result. The problem confronting the designer is that the adjustments all interfere with each other, making the problem miserably complicated.

The problem was first analysed by Ludwig von Seidel of Munich (1821–1896), who codified the five so-called *Seidel aberrations* of a monochromatic system and derived formulae for computing them. In the hands of subsequent optical designers, various theorems followed and ways were found of controlling, reducing and sometimes eliminating the aberrations. The theory which follows is therefore known as 'Seidel theory' and sometimes as 'first order theory' or 'third order theory'. It is not exact, but it gives a better solution than Gaussian optics, so that small adjustments can be made to give an acceptable design. The problem is that there are many variables and there may be many 'best' solutions such that a tiny variation in any of the variables gives a worse result; but perhaps the solution is only locally 'best'. An analogy is a

mountaineer climbing a peak in a mist. He may arrive at a summit, but is it the *real* summit of the peak or just the top of one of the foothills of the range?

B.2 Third order theory

Seidel was able to classify the aberrations into five species:

(1) spherical aberration,
(2) coma,
(3) astigmatism,
(4) field curvature and
(5) distortion.

They are not independent of each other and correcting one tends to exacerbate the others. Finding the balance is the art of optical design.

In addition to these five purely geometrical shortcomings there are two *chromatic aberrations* caused by the change in refractive index with wavelength: *longitudinal*, where the blue image may focus on a different surface from the red, and *lateral*, where the blue image is at the same focus but larger or smaller than the red. Both may be present at the same time of course. Despite Newton's belief in the irreducibility of chromatic aberration, a barrister, Chester Moore Hall (1703–1771) knew better and invented the achromatic doublet lens. The optician John Dollond, hearing indirectly of this invention, patented it but on learning the facts he did not enforce his patent. He was the first to make *and sell* achromatic doublet lenses, and his reputation for telescope making flourished.

B.3 The five Seidel aberrations

In what follows we refer to a narrow pencil of parallel rays as a *paraxial* bundle and to a set of parallel rays passing through the edges of the pupil [1] as the *marginal* rays. If the rays, although mutually parallel, are at an angle to the optic axis we refer to them as an *oblique* bundle. They were described briefly in Chapter 1. What follows now is a brief description of each of the aberrations.

B.3.1 Spherical aberration

This occurs when the marginal rays of a parallel axial bundle come to a focus at a different point from the focus of the paraxial ray bundle. The distance

[1] For most simple purposes the pupil is the rim of the lens or mirror.

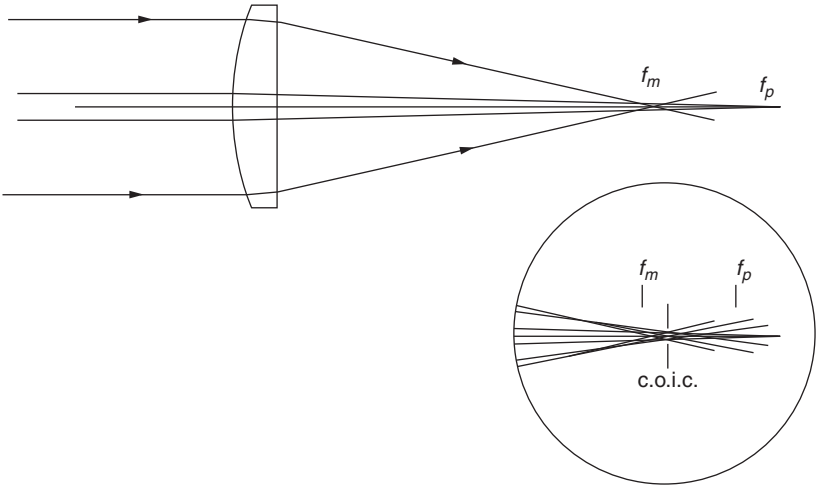

Figure B.1 Spherical aberration. A single lens with spherical surfaces, as shown here, always has positive spherical aberration, the marginal focal length being less than the paraxial focal length. In compound lenses it may well be zero or negative. In the simple instance the circle of least confusion is approximately one-third of the distance from the marginal focus to the paraxial focus.

between the paraxial focus and the marginal focus is the *longitudinal spherical aberration*.

B.3.2 Coma

This is when the marginal rays of an oblique bundle arrive at a focus different from that of the corresponding paraxial oblique bundle. In the absence of spherical aberration this marginal focus must be on the same plane as the paraxial focus. Two different types of coma must be distinguished. The marginal rays from the ends of a diameter in the same plane as the optic axis, known as the *meridional plane*, meet at the point of maximum *meridional* or *tangential* coma. The marginal rays from points on a diameter perpendicular to that plane meet at the point of *sagittal* coma. These two points are at the ends of a diameter of a small circle in the focal plane, and pairs of marginal rays from the ends of other diameters of the pupil meet at points on this circle.

B.3.3 Astigmatism

In the absence of coma and spherical aberration, the rays of an oblique bundle in the plane of the optic axis, which form a meridional image point, come to

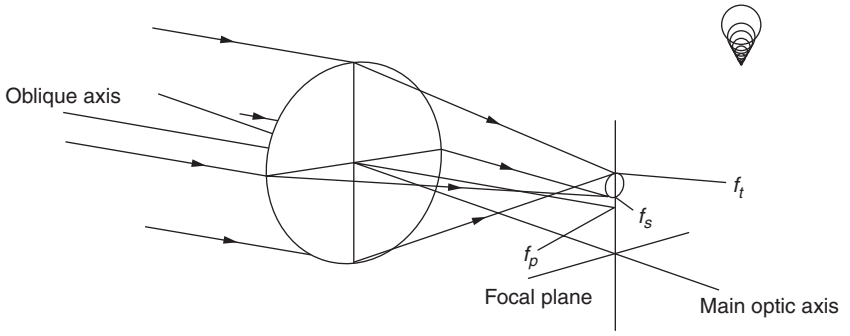

Figure B.2 Coma. Pairs of rays from opposite ends of any diameter will meet at a point on a circle in the focal plane. When the tangential focus is furthest from the optic axis, positive coma is indicated. Different annuli of the lens contribute different coma circles, and they lie nested between two lines inclined at 60°. The resulting aberration patch then resembles a comet tail, which gives the aberration its name.

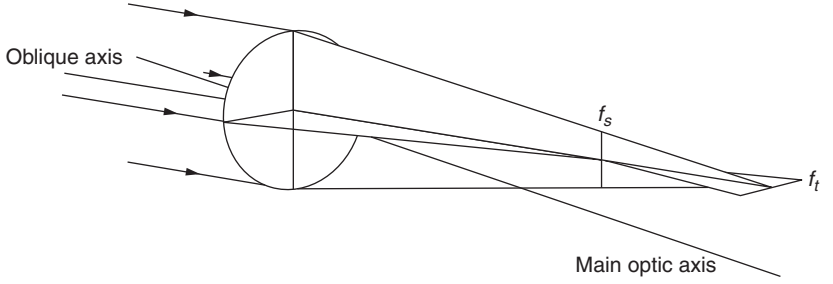

Figure B.3 Astigmatism. The same ray pairs which show coma are used. A lens with astigmatism would focus the rim of a wheel on a different surface (the tangential focal surface) from the spokes (the sagittal focal surface). Both surfaces are curved, and Joseph Petzval (1807–1891), who first investigated astigmatism, defined a third surface, the 'Petzval surface' which lies between them. It serves as a reference because its radius can be calculated and the astigmatism can be measured from this surface. The *Petzval sum* ($\mathrm{Pz} = \sum_1 / n_i f_i$), where n_i and f_i are the refractive index and focal length of the ith element in a compound lens or total instrument, measures this curvature, and provides the condition ($\mathrm{Pz} = 0$) that when the astigmatism is zero, the focal surface is flat. This is commonly known as the 'flat-field' condition.

a focus on a different focal plane from the rays forming the sagittal image point. The image of a point source at $-\infty$ is then a short line perpendicular to the optic axis at the meridional focus, and a similar short line lying in the meridional plane at the sagittal focus.

B.3.4 Field curvature

In the absence of the other aberrations, the focal surface of a centred system, where a distinct image of an extended object at $-\infty$ is formed, is spherical. In the presence of astigmatism two such spherical surfaces must be defined, corresponding to the sagittal and tangential foci. In fact a *third* surface, lying between the two, the *Petzval surface* is important. This is where the two constants **C** and **D** are important.

B.3.5 Distortion

This occurs when the position of the principal planes changes with obliquity, so that even though the focal surface is a plane, the magnification of the system changes with field angle. The image of a square object at $-\infty$ is then extended at the corners (pin-cushion distortion) or bulges at the sides (barrel distortion). Distortion can be tolerated (up to a point!) in a camera lens or a visual telescope but freedom from distortion is required in a lens to be used for survey work when critical measurements are to be made from the photographs. Even so, calibration exposures must be made and a correction table drawn up before field work can be analysed.

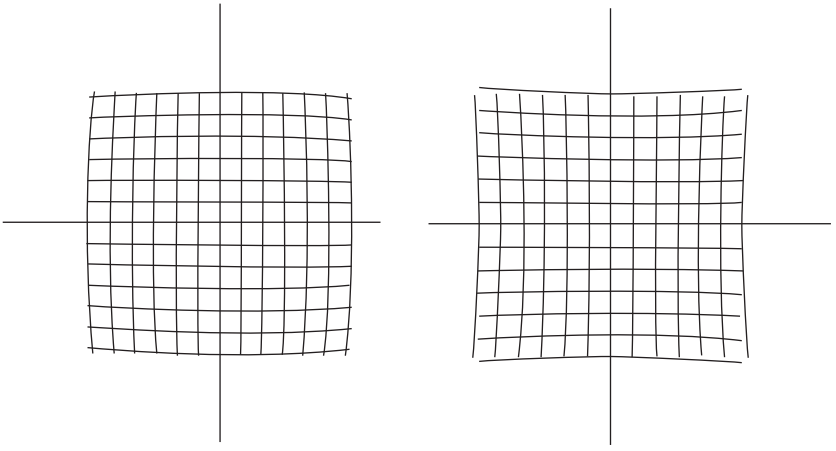

Figure B.4 Distortion. These are the images of a square net in a lens with barrel distortion or pincushion distortion. Wide-angle lenses in particular are apt to suffer barrel distortion. 'Symmetrical' lenses, i.e. those in which the second group of elements is the mirror image of the first, are to a good approximation free from distortion and coma, the 'odd' aberrations.

B.4 The Seidel formulae

Two quantities are initially useful. They are (1) the spherical aberration coefficient **B** and (2) the coma coefficient, $\mathbf{F_0}$. The *longitudinal* spherical aberration, Δ_{sph}, the distance from the paraxial focus to the marginal focus, is then *defined* by:

$$\Delta_{sph} = \frac{1}{2}\mathbf{B}y^2 f$$

where **B** is the spherical aberration coefficient, y the semi-aperture and f the focal length. For example, the longitudinal spherical aberration coefficient for a thin lens of focal length f and surface curvatures c_1 and c_2 is:

$$\mathbf{B} = c_1^2[n^3 - 2n^2 + 2]/n - c_1c_2[2n^2 - 2n - 1] + c_2^2n^2$$

Similarly, *tangential* or *meridional* coma, the distance from the point where the paraxial rays meet to the point where marginal rays *in the plane of the optic axis* meet, is described by:

$$\Delta_{coma} = 3\mathbf{F_0}y^2 \tan\alpha$$

where α is the inclination of the oblique chief-ray. The sagittal coma is one-third of this.

Again for a thin lens the coma coefficient is

$$\mathbf{F_0} = c_1[n^2 - n - 1]/n - nc_2$$

The suffix to the coma coefficient is important. This signifies the coma when the pupil of the lens is the aperture of the lens itself.

The corresponding quantities for a concave mirror are useful:

$$\mathbf{B} = (1 - e^2)c^2/2$$

and

$$\mathbf{F} = [c + \mathbf{Z}(1 - e^2)c^2]$$

where c is the curvature of the mirror, e the excentricity of the generating conic (1 for a paraboloid, 0 for a sphere) and **Z** the distance of the field stop from the mirror vertex.

If the pupil of the systen is at a distance **Z** from the lens or mirror, *Fraunhofer's theorem* shows that the coma coefficient becomes

$$\mathbf{F} = (\mathbf{F_0} + \mathbf{ZB})\tan\alpha$$

This is interesting because it means that if the stop can be placed at a distance $-\mathbf{F_0}/\mathbf{B}$ from the lens, the coma is eliminated – at least in the Seidel limit. The minus sign implies that the stop is to the left of the lens, although it is quite usual for a system to have negative coma. Notice also that you cannot eliminate coma in this way if there is no spherical aberration – the coma of a paraboloidal mirror is the same as that of a spherical mirror. The spherical aberration coefficient of a spherical mirror is $1/8f^2$ so that the position for a stop to eliminate coma is at $Z = -2f$, which is the centre of curvature of the mirror. This is where the corrector plate of a Schmidt camera is placed. Coma correctors for astronomical telescopes depend on introducing a trivial amount of spherical aberration and relying on the fact that the pupil is the objective itself and the pupil distance $\mathbf{Z}$ to the corrector is very large, so that zero power and small (below the diffraction limit) spherical aberration conspire with Fraunhofer's theorem to give a comparatively large coma-free field.

B.4.1 Coma correction with a stop

A simple example gives the geometrical flavour of these formulae. Consider a lens in the form of a hemisphere with the plane face on the object side. If a stop is placed centrally on the plane face, then collimated light (i.e. coming from a point source at $-\infty$) will be refracted and transmitted as a parallel beam until it reaches the curved surface. There is no preferred optic axis and the beam will converge symmetrically to a focus on a spherical focal surface concentric with

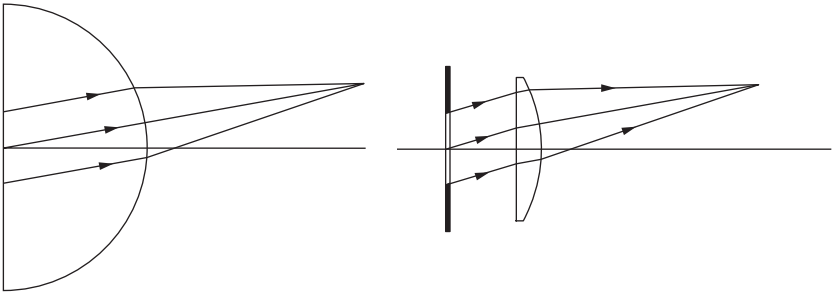

Figure B.5 An example of aberration correction with a stop. A hemispherical lens has no preferred optic axis. Any line passing through the centre of the sphere will serve. Any parallel ray-bundle is symmetric if it passes through a stop symmetric about the centre and therefore has no coma. Most of the lens can be removed to leave an ordinary plano-convex lens, and the image of the stop in the lens will ensure that the rays *appear* to come from the centre and again are symmetric and therefore have no coma.

the surface of the lens. The image point will have only spherical aberration (and of course field-curvature).

Now, put the plane surface of the lens as close as possible to the spherical surface and put a stop at the point on the optic axis (there is now an axis!) a distance $-R/n$ from the plane vertex. Collimated light coming through the stop will reach the curved surface of the lens at the same angle as the corresponding bundle in the hemispherical lens, and again there will be no coma. Any other position of the stop will introduce an asymmetry and coma will appear. The stop position is critical.

These formulae for the coefficients are sometimes found with lens radii instead of curvatures. In practice the curvatures are more convenient, partly because the focal lengths are not involved and partly because it is curvatures which appear in ray-tracing software. In the case of plano-convex lenses, where either c_1 or c_2 is zero, the formulae simplify considerably, making quality assessment a simple matter.

Two more coefficients, **C** and **D**, are useful as they allow the astigmatism and field curvature to be treated. The two aberrations are so closely interlinked that they are generally treated together.

B.4.2 Petzval's theorem

Joseph Maximilian Petzval (1807–1891), a European[2] mathematician at the University of Budapest, discovered a relationship between the curvatures of the meridional and sagittal focal surfaces. He introduced the *Petzval surface* and defined its curvature as the *Petzval sum* given by:

$$C_p = \sum_i \frac{1}{n_i f_i}$$

where n_i and f_i are the refractive index and focal length of the ith element of a centred optical system.

The meridional surface is defined as a sphere with vertex at the paraxial focus and curvature given by

$$C_m = 4\mathbf{C} + 2\mathbf{D}$$

and the sagittal surface, likewise has a curvcature

$$C_s = 2\mathbf{D}$$

[2] Of uncertain and much contested ethnicity.

and then Petzval's theorem states that:

$$\mathbf{C} - \mathbf{D} = \sum_i \frac{1}{n_i f_i}$$

The sagittal and meridional surface curvatures are then:

$$C_m = C_p + (3/2)\Delta C$$

$$C_s = C_p - (1/2)\Delta C$$

For a mirror, spherical or parabolic, the coefficients are:

$$\mathbf{C} = \frac{1}{2f} + \frac{\mathbf{Z}}{2f^2} + \frac{\mathbf{Z}^2(1-e^2)}{8f^3}$$

$$\mathbf{D} = \frac{\mathbf{Z}}{2f^2} + \frac{\mathbf{Z}^2(1-e^2)}{8f^3}$$

Astigmatism and field curvature can be tolerated to some extent in the design of eyepieces for example, where the criteria are freedom from chromatic aberration and a wide field of view. In a camera lens a flat field is the overriding consideration and both meridional and sagittal fields must be as flat as the designer can make them.

B.5 The aplanatic lens

Originally the word *aplanatic* was intended to mean 'free from spherical aberration', but currently it means 'free from spherical aberration and coma'. Aplanatic lenses may still have astigmatism but there is one remarkable arrangement which is free from all of these.

Consider a glass sphere. Take any line through the centre as the optic axis. Direct a ray from any exterior point towards the point on this axis which is at a distance $R(1+n)$ from the surface (the vertex). Then using the law of sines it is easy to show that the ray will be refracted to arrive at a point on the same axis at a distance $R(1+1/n)$ from the vertex. This is exactly true for a ray from *any* direction: there is no Gaussian or Seidel approximation.

There is no need to use the whole glass sphere. A second spherical surface centred on the arrival point will cause no further refraction and the resulting meniscus lens is called an aplanat. It can be shown that a small image area around this focus is also, for all practical purposes, aplanatic. The drawback is that the lens has a virtual object (at the $R(1+n)$ point) and so cannot be used

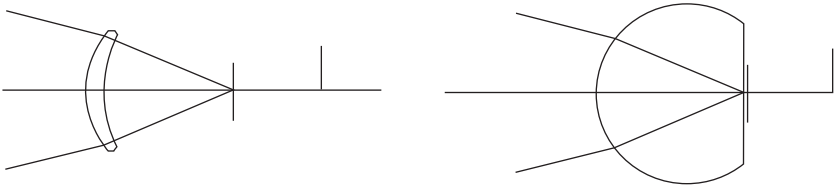

Figure B.6 An aplanatic lens. A consequence of Euclidean geometry is that a ray directed to one particular point on the axis $Z = R(1 + n)$ from any direction will cross the axis at $Z = R(1 + n)/n$. This makes the point a perfect image of the virtual point object. A small area around that image is also free from spherical aberration and coma, as required by the term *aplanatic*. A semiconductor detector has noise power proportional to its area and immersion in a transparent aplanat allows this area to be reduced by a factor n^2.

alone. Its chief use is to increase by a factor n the focal ratio of the system of which it is part. This can help in one of two ways.

(1) It can reduce the focal ratio so that a convergent light beam can be focused on to a smaller detector. Bearing in mind that the noise-equivalent power of a detector increases as the square-root of its sensitive area, this has substantial advantages when observing weak signals. Some infra-red detectors are sold already embedded in a plastic aplanat.
(2) It increases the focal ratio in a microscope objective, and can do this several times if several aplanats are used in series. In practice chromatic aberrations are corrected by lens combinations which together are approximately aplanatic.

B.6 Chromatic aberrations

These are consequences of the dispersion of light by a dense medium such as glass.

B.6.1 Longitudinal colour

Paraxial rays of blue light, being more refracted than red, may come to a focus closer to the P_2 principal plane, and images are seen either red with a blue border or blue with a red border, depending on where the focal surface is placed. It was this aberration, *longitudinal colour*, which was first analysed by John Dollond while investigating chromatic dispersion in different glass types, and was the discovery which led to the invention of the *achromatic doublet* lens.

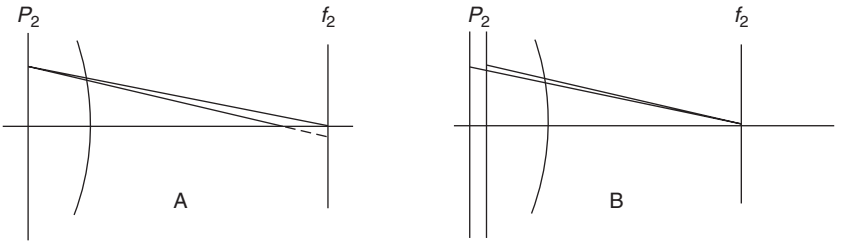

Figure B.7 Chromatic aberration. There are two conditions which may cause an image to have colour fringes. Either (1) the focal length is different for different wavelengths so that there is no common focal plane, or (2) all rays focus properly at the same focal plane but the principal planes are different for different wavelengths. This means that the magnification of the image is different and, for example, the red image may be larger than the blue, giving a red fringe to any dark object. This is called 'lateral colour'.

B.6.2 Lateral colour

Although the red and blue images may come to a focus on the same focal plane, the principal planes and consequently the focal length may be different for red and blue light. The blue image then will be slightly larger or smaller than the red image and there will be the same kind of colour-fringed image at the focus.

B.7 Diffraction

This is the consequence of the wave nature of light. An image is formed by wavefronts originating in the object and arriving at the geometric image point with different phases, so that there is interference when phases are opposed and reinforcement when the phases are the same. The consequence is that an image 'point' may be criss-crossed with very fine interference fringes. There is thus a limit to the detail that can be recognized in the image of a point source. At best a point source is imaged as an *Airy disc* of diameter, Δ, given by the *Airy formula*

$$\Delta = 1.22F\lambda$$

where F is the focal ratio of the final image-forming aperture and λ the wavelength of the light. The disc is surrounded by a faint ring of light which contains 7.2% of the total power. When accurate photometry is being done this is bright enough to obscure the presence of a faint companion to a star for example.

In photographic practice a higher F-number will reduce the geometrical aberrations so that a camera gives a progressively sharper image as the iris is closed down; but then the diffraction disc becomes larger. Consequently there is an optimum F-number which will give the sharpest image. It depends very much on the particular lens type but in general it will lie somewhere between F/6.3 and F/11.

Appendix C

A brief introduction to Fourier optics

Fourier transforms provide an important and simple way of solving problems in diffraction theory and the theory of optical imaging.

In 1807 J.-B. J. Fourier (1768–1830) discovered that there exists an unexpected and remarkable reciprocity between two related functions. A *Fourier pair* of functions $f(x)$ and $\phi(p)$ can be defined by:

$$\phi(p) = \int_{-\infty}^{\infty} f(x) e^{2\pi i p x} dx \tag{C.1}$$

and

$$f(x) = \int_{-\infty}^{\infty} \phi(p) e^{-2\pi i p x} dp \tag{C.2}$$

where the variables x and p are called **conjugate variables** and the functions $f(x)$ and $\phi(p)$ are a **Fourier pair**. [1] They are usually written

$$\phi(p) \rightleftharpoons f(x)$$

Any analytic function $f(x)$ can have a Fourier pair provided

(1) that it is single valued,
(2) that is is square integrable, which is to say that the ingetral of $[f(x)]^2$ from $-\infty$ to ∞ is finite,
(3) that it is piecewise continuous, and
(4) that it has upper and lower bounds. [2]

[1] It is important that the exponent in one equation but not the other has a negative sign. To that extent the reciprocity is broken.

[2] This last condition is viewed with suspicion by physicists in view of the fact that the Dirac delta-function, which has no upper bound, works well and has unity as its Fourier pair.

At first, Fourier analysis and its theorems proved to be a very useful tool in pure and applied mathematics but soon it was found that the abstract variables x and p could be given physical meanings, particularly in *spectrum analysis* where one of the variables could represent *time* and the other *frequency*. A time-varying quantity, $f(t)$, when Fourier transformed, gave its frequency content, $\phi(\nu)$.

Fourier transforms also found an application in physical optics. In Fraunhofer diffraction, the amplitude diffracted at an aperture is calculated by integrating the infinitesimal amplitudes of all the Huygens wavelets in the aperture, taking due account of their relative phases, and this integral turns out to be the standard Fourier integral, with x as the position of the wavelet in the aperture and $f(x)$ its amplitude. Its conjugate variable, p, is $\sin\theta/\lambda$, where θ is the diffraction angle and λ the wavelength of the light. It all leads to the simple theorem:

> the amplitude distribution on the focal plane of a lens is the Fourier transform of the amplitude distribution on the lens aperture.

The familiar problem of diffraction by a single slit, where all the Huygens wavelets have the same amplitude A from $x = -a/2$ to $x = a/2$ and zero everywhere else, can be solved easily. The light coming through the slit can be represented by a 'top-hat' function, $\Pi_a(x)$, of width a and unit height. In this case the amplitude is A, so that the integral to be done is

$$\int_{-a/2}^{a/2} A.e^{2\pi ipx}dx$$

which is the Fourier integral, but with the infinite limits replaced by the slit ends since there is no contribution to the integral outside the slit.

The light diffracted by the slit then has an amplitude:

$$\phi(\theta) = Aa.\mathrm{sinc}(\pi a \sin\theta/\lambda) \tag{C.3}$$

where $\mathrm{sinc}(x)$ is the 'sinc-function', $(\sin x/x)$, sometimes called the 'sine-cardinal'. (The sinc-function, and its Fourier pair the 'top-hat' function, are a most important pair of functions in physics.)

The *intensity* of the diffracted light, a function of θ, is the square modulus [3] of this and is the familiar 'sinc-squared' curve of single-slit diffraction theory. Notice that the first zeros of this pattern, on either side of the central maximum,

[3] The diffracted *amplitude* involves the phase-factor and is probably complex. The intensity must of course be real – it can be measured!

are diffracted at an angle θ given by $\sin\theta = \lambda/2a$, so that θ increases and the pattern spreads as the slit width, a, narrows.

With this in mind, the diffracted amplitude of any aperture can be calculated, with a grid of lines for example, represented by a square-wave. The idea extends to two dimensions and the two-dimensional Fourier transform can be converted to polar coordinates, when it becomes a *Hankel* transform. It then can be used to describe diffraction from a circular or annular aperture such as a camera lens or a Cassegrain telescope objective.

Although it was Kelvin who first pointed out the application of Fourier analysis in Fraunhofer diffraction, it appears to have been forgotten (at least at undergraduate level) until the 1950s when Henry Lipson, in the Manchester Institute of Science and Technology,[4] applied the method in his lectures on physical optics and later in a seminal undergraduate optics textbook. The advent of the so-called 'fast Fourier transform' (FFT) in computer science greatly simplified calculation and did much to revolutionize diffraction studies.

C.1 Convolutions

One of the more important concepts which are dealt with by Fourier theory is the notion of the *convolution* of two functions. The standard one-dimensional 'hand-waving' (i.e. non-mathematical) description of a convolution comes from spectroscopy. A purely monochromatic spectrum line in a 'perfect' spectrometer would be an infinitely sharp image.[5] In practice spectrometers are not perfect and the graph of intensity against wavelength follows a curve, probably like a Gaussian curve, thin and high and with the area under it representing the total radiated power passing through the spectrometer. This curve is called the *instrumental profile* and is denoted usually by $I(\lambda)$, which means that the spectral intensity curve will be $I(\lambda - \lambda_0)$ when there is unit power in a spectrum line at λ_0.

The graph of a continuous spectrum has power-per-unit-wavelength range versus wavelength, and can be representeed by a function $S(\lambda)$ when it enters the spectrometer; and the question is, what is the apparent power-per-unit-wavelength measured at the output?

The answer is that it is the *convolution* of $S(\lambda)$ with $I(\lambda)$. Each infinitesimal element of the original spectrum is replaced by an infinitesimal $I(\lambda)$ curve of the appropriate height, so that we replace the element at (λ_0) of height $S(\lambda_0)$ containing power $S(\lambda_0)d\lambda_0$, with the curve $S(\lambda_0)d\lambda_0.I(\lambda - \lambda_0)$. The

[4] Now subsumed into the University of Manchester.
[5] A Dirac delta-function.

first factor is the power by which the second (unit power) instrumental profile must be multiplied.

Next, to find the total power apparently at wavelength λ due to all elements of the spectrum we integrate and find that

$$S_{out}(\lambda) = \int_{\infty}^{\infty} S_{in}(\lambda_0).I(\lambda - \lambda_0).d\lambda_0$$

The integral on the right-hand side is the convolution of $S_{in}(\lambda)$ with $I(\lambda)$ and is usually written

$$C(\lambda) = I(\lambda) * S(\lambda)$$

It can also be written

$$C(\lambda) = S(\lambda) * I(\lambda)$$

because the convolution process is commutative.

The idea extends to two or more dimensions and is important in photography, where the image produced at the focal surface is the convolution of the distant object with the *point spread function* of the camera lens, the image the lens would produce of a point object at $-\infty$.

The importance of convolutions in Fourier theory rests on the remarkable *convolution theorem*, which states that the Fourier transform $\overline{C}(s)$ of the convolution $C(\lambda)$ is the *product* of the Fourier transforms of its constituents, $\overline{I}(s).\overline{S}(s)$. Here s is clearly the Fourier conjugate to λ.

This theorem is of profound importance in Fourier theory generally and in its application to diffraction optics. Using the fast Fourier transform algorithm, the computer process: transform$_1$–transform$_2$–multiply them–do the inverse transform of product, is wonderfully less complicated than doing a point-by-point convolution of two functions.

C.2 The diffraction grating

One of the more useful functions developed especially for physicists is the *Dirac comb*, named (without permission) after P. A. M. Dirac (1902–1984) of Cambridge University, the inventor of (among other things) the Dirac delta-function $\delta(x)$, the peculiar function with the properties (a) that it is zero everywhere except at $x = 0$, where it is infinite, and (b) that its integral from $-\infty$ to ∞ is unity. The Dirac comb is an infinite set of equally-spaced delta-functions, denoted by the Cyrillic letter Ш (shah) and has the remarkable property that its Fourier transform is another Dirac comb:

$$\text{Ш}(x) \rightleftharpoons \frac{1}{p}\text{Ш}(p)$$

With this we are in a position to describe simply the working of a diffraction grating.

The *aperture function* of a grating is the convolution of a narrow top-hat function – the aperture function of a single slit – and a Dirac comb.[6] This gives an infinite set of equally spaced top-hat functions, or a 'square-wave' as it is called in electronics. Because the width of a grating is finite the whole thing must be multiplied by a very broad top-hat function representing the whole width of the ruled area. Now we Fourier transform this whole aperture function to get the angular spread of amplitude in the diffracted beam. The convolution becomes a product and the product becomes a convolution. The convolution first:

$$\Pi(x) * \text{Ш}(x) \rightleftharpoons \frac{1}{p}\text{sinc}(p).\text{Ш}(p)$$

where $\text{Ш}(p)$ represents the Dirac comb and $\text{sinc}(p)$ is the Fourier transform of $\Pi(x)$, the position x being measured across the grating. The convolution of the right-hand side with the Fourier transform of the very wide top-hat simply replaces each tooth of the Dirac comb with the very narrow sinc-function representing the diffraction by the whole width of the grating. This narrow sinc-function is the way each monochromatic spectrum line is displayed. The new Dirac comb now gives the *orders* of diffraction, and the variable p is as before, $\sin\theta/\lambda$ with θ the angle of diffraction.

C.3 Fourier transform infra-red spectroscopy

The idea of using a Michelson interferometer to do spectroscopy originated in the 1950s and was slow to develop because of the extraordinary technical difficulties encountered. There were three reasons why it was pursued.

(1) The light throughput (étendue) of an interferometer is about 300 times greater than that of a spectrometer with a similar pupil diameter, basically because the entry slit is replaced by a circular aperture.
(2) If the main source of system noise is in the detector rather than photoelectron shot-noise from the radiation, there is a so-called *multiplex Advantage* because radiation from all wavelengths is fed to a single detector,[7] and

[6] Convolution here simply means replacing each tooth of the comb with the narrow top-hat function.
[7] Rather than to N detectors as in a spectrograph.

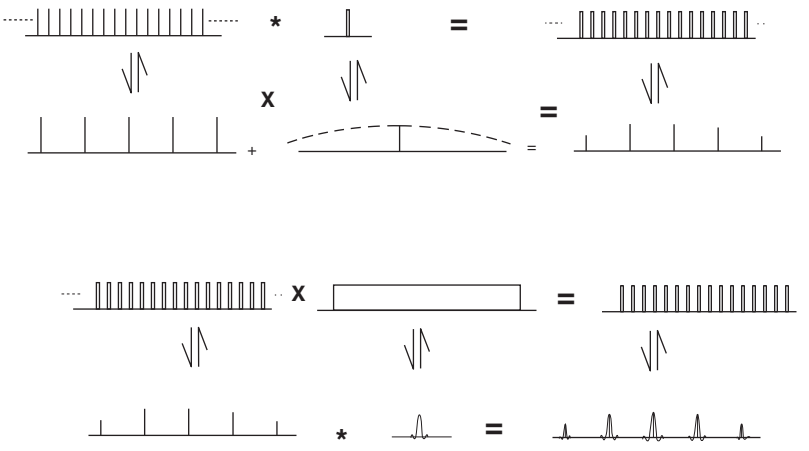

Figure C.1 The Fourier theory of a transmission grating. The fundamental theorem is that in an image-forming system, the light-amplitude distribution in the focal plane is the Fourier transform of the light-amplitude distribution in the aperture. In a diffraction grating, an individual ruling is represented by a narrow top-hat function and the whole grating width is represented by a broad top-hat. The *convolution* of the narrow top-hat with a Dirac comb gives an infinite array of equally spaced top-hat functions and this is curtailed when it is multiplied by the broad top-hat. This now represents the aperture of the grating. The rule in doing Fourier transforms is that on transforming, a product becomes a convolution and a convolution becomes a product. Thus the broad top-hat becomes a narrow sinc-function, the dense Dirac comb becomes a sparse Dirac comb and the narrow top-hat becomes a very wide sinc-function. The net result is a widely spaced comb of narrow sinc-functions under a very wide sinc-function. Remember that after the Fourier transform the dimension is an angle, and the resultant function represents the different orders of diffraction, each modified in amplitude by the broad sinc-function.

the signal-to-noise ratio in the output signal is enhanced by a factor $\sqrt{N}$, where N is the number of resolved elements in the spectrum.

(3) The advent of the computer algorithm known as the *fast Fourier transform* in 1964 made it possible to Fourier transform long strings of data – anything up to 10^6 words – in a few seconds or a few minutes, and this made a realistic spectral resolution possible.

These three factors together promised a major advance in infra-red spectroscopy and this was fulfilled towards the end of the twentieth century, with commercially made FTIR spectrometers becoming widely available. On the other hand, in the visible and ultra-violet regions, where photoelectron shot-noise predominated, the multiplex advantage not only disappeared, but a

large *dis*advantage appeared, since the shot-noise on a strong emission line is distributed throughout the spectrum by the Fourier transform process and swamps the faint signals from any weaker lines.

The principle of FTIR spectroscopy is as follows.

- A monochromatic spectrum line, shining through a Michelson interferometer, will give a signal in the detector which varies sinusoidally as the path-difference is steadily increased. The *period* of the signal depends on the wavelength of the light.
- The Fourier transform of a sinusoid is a Dirac delta-function of height proportional to the sinusoid amplitude, and at the frequency of the sinusoid in the spectrum.
- When a whole spectrum is transmitted, each component will add a sinusoid with its own intensity and period to the resultant net signal and this totality is recorded as a function of time as a data string in a computer memory.
- The Fourier transform of this data string is the spectrum of the received signal.

The result was that an infra-red spectrum which previously had taken a quarter of an hour to measure, could now be measured in seconds, and FTIR spectrometers could be used 'on-line' to control and monitor industrial chemical processes.

The technical difficulties which were overcome were (a) that of maintaining interferometric alignment to better than $\lambda/20$ during the path-changing process, and (b) since the data collection process involved stop-and-measure at each sample point, ensuring that each step was the same length to the same $\lambda/20$ degree of accuracy. In effect, a FTIR spectrometer is equivalent to a grating-ruling engine, except, that it (very quickly) rules a new grating each time a spectrum is measured.

Further reading

Many of the books listed below are from an earlier era, sometimes more than 50 years past. There appear to be few modern texts on geometrical or physical optics which have received universal acclaim, and there is much to be gained by the alternative approaches of earlier, gifted authors. Your present author recommends the Lipsons' book particularly for their clear explanation of the use of Fourier transforms in physical optics, and Kingslake's book on optical system design. In optical physics (the interaction between light and matter), along with Lipson I would shortlist Ditchburn, Garbuny and Strong which, although ancient scripts, are clearly written, readable and exact and in no sense out of date: laser physics alone is missing from their texts. For geometrical optics alone, Welford is good at undergraduate level.

Born and Wolf is the physical optics 'bible', the final court of appeal so to speak, but it is not for the young or the beginner. Likewise, Conrady is the prime authority on geometrical optics and optical design, but again this is for the library shelf rather than the private student bookshelf.

Born, M. and Wolf, E., *Principles of Optics* (Cambridge University Press, Cambridge, 7th edition, 1999).

Bracewell, R. N., *The Fourier Transform and its Applications* (McGraw-Hill, New York, 1965).

Brigham, E. O., *The Fast Fourier Transform* (Prentice Hall, New York, 1974).

Clarke, D. and Grainger, J. F., *Polarized Light and Optical Measrement* (Pergamon, Oxford, 1971).

Conrady, A. I., *Applied Optics and Optical Design* (Dover, New York, 1960).

Ditchburn, R. W., *Light* (Blackie, London, 1952; Dover, New York, 1991).

Garbuny, M., *Optical Physics* (Academic Press, New York, 1965).

Harrison, G. R., Lord, R. C. and Loofbourow, J. R., *Practical Spectroscopy* (Prentice Hall, New York, 1948).

James, J. F., *Spectrograph Design Fundamentals* (Cambridge University Press, Cambridge, 2007).

James, J. F., *A Student's Guide to Fourier Transforms* (Cambridge University Press, Cambridge, 2011).
James, J. F. and Sternberg, R. S., *The Design of Optical Spectrometers* (Chapman and Hall, London, 1969).
Kingslake, R., *Lens Design Fundamentals* (Academic Press, New York, 1978).
Kingslake, R., *Optical System Design* (Academic Press, New York, 1983).
Kingslake, R., *A History of the Photographic Lens* (Academic Press, San Diego CA, 1989).
Lipson, A., Lipson, S. G. and Lipson, H., *Optical Physics* (Cambridge University Press, Cambridge, 2011).
Malacara, D. (editor), *Geometrical and Instrumental Optics*. (Academic Press, New York, 1988).
Malacara, D. and Malacara, Z., *A Handbook of Optical Design*. (Marcel Dekker, New York, 2004).
Minnaert, M., *Light and Colour in the Open Air* (Dover, New York, 1954, 2012).
Saleh, E. A. and Teich, C., *Fundamentals of Photonics* (John Wiley & Sons, New York, 1991).
Smith, W. J., *Modern Optical Engineering* (McGraw-Hill, New York, 3rd edition, 2000).
Strong, J., *Concepts of Classical Optics* (W H Freeman, San Francisco, CA, 1958).
Walker, S. and Straw, H., *Spectroscopy* (Chapman & Hall, London, 1962).
Welford, W. T., *Geometrical Optics* (North-Holland, Amsterdam, 1962).
Wood, R. W., *Physical Optics* (Macmillan, New York, 1905, 1942).

Index

aberrations 13
  chromatic 14, 171
    lateral 14, 163
    longitudinal 14, 163
  geometric, 50, 75, 90, 92
  off-axis 14, 26, 30
  optical 162 *et seq*.
  Seidel 162
  spherical 13, 163
albedo 41
alignment 80, 92
aperture 6
  numerical 7
  relative 6

Bayer, Bryce 59
Bayer mask for CCD camera 59
black and white grids 54, 59
black colloidal silver 57
black surface 41
black body radiation 138 *et seq*.
black body spectrum 78
bolometer 132 *et seq*.
Brewster, Sir David 106
Brewster angle 106, 107

camera 43
  CCD 58
    cooled 61
  Leica 49, 59
  pinhole 47
  rising-front 51
  Schmidt 25
  single lens reflex 46
  tilt-and-shift 51
camera focusing 50
camera lenses 49

CCD
  calibration 62
  camera 57
    lens, radioactive 62
  cooled 60
chief-ray 10
cleaning 141
  camera lenses 141
  of diffraction gratings 143
  of mirrors 143
  optical surfaces 141
  plastic lenses 143
coelostat 124
  Lippmann 126
  two-mirror 126
coherence 54, 85
coherence area 85
coherence length 86
collimator 11, 160
colour 30
  eyes and 32
  in nature 40, 41
  lateral 27, 172
  non-spectral 41
  perception of 39
  primary 40
  pure 40, 41
  spectral 40
colour blindness 40
colour film 59
colour fringing 14
colour photography 57
colour printing 40, 42
colour sensitivity 39
colour television 42
colour triangle 40, 41

colour vision 39
conjugate points 10
convolutions 176 *et seq*.
cosmic ray events 61
cross-hairs 14, 145

depth of focus 51
diffraction 4, 13, 54
  Fraunhofer 72, 175
diffraction grating 69
  concave 75
  ghosts 80
diffraction spot 50
  order of 70
  order-sorter 70
Dirac, P. A. M. , 177
Dirac comb 177 *et seq*.
Dirac delta-function 177, 180
distortion 14, 32 *et seq*., 45, 163, 165

emissivity 140
  function 138, 140
étendue 8
  conservation of 8
eye38
  accommodation of 30
  acuity of 38
  colour perception 39
  vibrations of 38
eyepieces 30
  Abbé orthoscopic 33
  Erfle
  Huygenian 31
  Kellner 33
  Plössl 34
  Ramsden 33
eyepiece graticules 36, 37
eyepiece properties 32
eye-relief 16
  rifle sight 18

Faraday effect 110
field 11
  angle 47
  curvature 14, 163
  stop 9,
filters
  Christiansen 147
  colour 40, 41
  dichroic 147
  didymium glass 147
  dye 95, 145
  high pass 146
  interference 10, 146
  notch 146
  optical 145
  order-sorting 70, 147
fish-eye lens 47
flexible image-transport system 60
focal length 2
  negative 45, 46
focal point 12, 44
focal plane 10, 12
  stop 22
focal ratio 6
focal reducer 27, 28
focal surface 3, 14, 44
Foucault knife-edge test 51
Foucault method 82
Fourier, J.-B. J. , 174
Fourier optics, 174 *et seq*.
Fourier pair 174
Fourier transform 174 *et seq*.
Fraunhofer diffraction 175
Fraunhofer lines 66, 83, 155
Fraunhofer theorem 167

Gauss, C. F. 43
Gaussian achromat 43
Gaussian optics 11 *et seq*.
  approximation 12, 154
graticules 36, 37

Helmholtz–Lagrange invariant 8

image
  collimated 27
  erecting prisms 19
  inverters 29
  real 2, 15, 16, 18, 31, 53, 151
  rotators 27, 28
  virtual 15, 17, 31, 86, 87, 151

interferometers 84
  Fabry–Pérot 93, 95
  Fizeau 54, 88, 90
  Jamin 100
  Mach–Zehnder 99
  Michelson 91, 92
  Sagnac 101
    cat's eye 92
    cube-corner 92
    stellar 88
    tilting 93
  Twyman-Green 88

Kerr effect 111

lens 2
  achromatic doublet 43
    air-spaced 142
  aplanatic 170
  Boys points of 55
  Cooke triplet 45
  double Gauss 44
  eye 16, 17
  field 16, 35, 160
  fish-eye 47
  focal ratio of 6, 7
  Fresnel 118
  lighthouse 119
  objective 15, 16 *et seq*.
  plano-convex 2
  point-spread function of 177
  radioactive 62
  rapid rectilinear 45, 52
  telephoto 45
  Tessar 41, 45
  unmounted 141
  vari-focal 49
  zoom 49
lens aperture 6, 7
lens pupil 6

matrix
  displacement 159
  ray transfer, 157 *et seq*.
microscopes 119
  Abbé theory 120
  compound 120
  Köhler illumination 120
  objective 121
    oil immersion 155
  travelling 122
mirror 3
  concave 4, 22
  infra-red 140
  objective
  primary 23, 25,
  secondary 23, 24
  spherical 26
mirror image 14, 28, 29
monochromators 71 *et seq*.

night glass 20

object
  virtual 170, 171
objective
  achromatic 142, 145
  astrographic 145
  binocular 19, 20
  immersion 7
  microscope 120
    oil-immersion 155
  telescope 15
objective lens 7, 10
optic axis, defined 1,
optical activity 107, 110
  induced 110
optical resonant cavity 160
optical screw 114
optical transfer function 54
optics
  Fourier 174
  Gaussian 11, 149
  geometrical 162, 181

Peltier cooler 60, 135
photoconductive effect 135
photoelectric effect 135
photoelectron 133
  shot-noise 133
photography
  architectural 51
  astronomical 27, 35
  colour 39
  digital 60
  high resolution 50
  landscape 109
  micro 49, 59
  Schlieren 51
  silver halide 57
  with a polarizing filter 109
photomultiplier 132, 135
photon 86, 132
  counter 135
  emission rate 133 140
photovoltaic effect 135
Planck, Max
  black body radiation curve 139
  radiation formula 139, 140
plate
  corrector 26, 27, 168
  étalon 99
  face 144
  half-wave 108
  photographic 26, 76
  quarter-wave 108, 110
  quartz 93
  wave 112, 113
plate camera 51, 52

plate measuring
machine 123
Pockels effect 113
point-spread function 54, 177
polarization 106
circular 105
plane of 104
polarization effects 105
prism,
Dove 29
erecting 19
Porro 19
roof, Amici 20
rotating 28
Schmidt–Pechan 19, 29
spectroscope 64
Purkinje effect 39

ray 3,
chief 10
reflected 23
ray bundle 10
ray diagram 4
ray path 5, 12
geometrical 5
optical 5
reduced 6
ray-tracing 12, 44
Rayleigh criterion 65
resolving power 72
Rayleigh refractometer 87
refraction 149 *et seq*.
law of sines 4
refractometer
Rayleigh 88
resonator
optical 103, 158
Ronchi grid 53
Rowland, H. A. 75
Rowland circle 76
Rowland ghosts 80

Scheimpflug principle 51
Schlieren photography 51
system 100, 101
Schmidt, Bernhardt, 25
Schmidt camera 27, 168
Schmidt stop 77
screen
fluorescent 137
projection 118
scattering 136
television 39, 118

spectra
continuous 64
fluted 64
types of 137
spectral colours 40, 41
spectral pass-band 10,
spectrograph 71
alignment 80
concave grating 75, 76
Čzerny–Turner 74
Ebert–Fastie 73, 74
Littrow 67, 68
spectrography 60
spectrometer 64 *et seq*.
multiplex Fourier 91
spectroscope 40
direct vision 69
prism 64 *et seq*.
spectroscopy 69
double-beam 79
Fourier transform 178
high resolution, 97
infra-red 178
interference 89, 97
Raman 80
ultra-violet
vacuum, 76
Stefan–Boltzmann law 138
stop 6
antiglare 6, 18,
coma-correcting 168
Schmidt 77
telecentric 10, 22
Strehl ratio 54

telecentricity 10
telescope 15
astronomical 48
binocular 16, 19
Cassegrain 23
reflecting 22
erecting 17, 18
Galilean 16, 17
Gregorian 23, 25
Keplerian 17
Maksutov 24
Newtonian 23
refracting 16
rifle sight 18
Ritchey–Chrētien 24
Schmidt 25
visual focusing of 30

television
  colour 47
television screen 118
theodolite 21
tiltmeter 103
ultra-violet
  far 136
  vacuum 137
  transparency 137
UVOIR region 134
Verdet's constant 110
vignette 9
visibility, fringe 54
vision
  colour 39
  night 39
  photopic 38
  scotopic 38, 39
wavefront 26, 51, 70
  division of 86
white
  synthetic 40
white light 40, 41, 42
white surface 41
Wien's displacement law 138
Young, Thomas 39
Young's two-slit experiment 86, 87
zoom lenses 49
zoom ratio 49